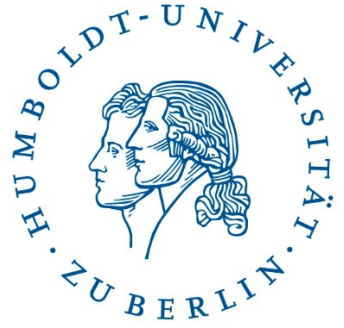

Never Trust the Teller! Feedback Manipulation and its Impact on Perceptual Inference

DISSERTATION

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
im Fach Psychologie

eingereicht an der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von Rekha Sreekumar Varrier, M.Sc Cognitive and Clinical Neuroscience

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät:
Prof. Dr. Bernhard Grimm

Gutachter/Innen:
1. Prof. Dr. phil. Martin Rolfs
2. Prof. Dr. med. Florian Schlagenhauf
3. Prof. Dr. med. Philipp Sterzer

Tag der Verteidigung: 19.02.2020

Statement of Authorship

I was the primary researcher in both of the studies described in this monograph. I was responsible for the major aspects of the experiments including programming the experimental tasks, conducting the experiments, analysing the data and preparing the studies for peer-reviewed publications. The research goals were set together with my thesis supervisor Prof. Philipp Sterzer, and the experimental design and analyses were finalised by us together with PD Dr. Matthias Guggenmos. The simulation described in Chapter 2 was performed by Dr. Heiner Stuke.

Funding

This thesis was funded by the German Research Foundation (DFG)-funded Research Training Group GRK 1589/2 of the Bernstein Center for Computational Neuroscience (BCCN) Berlin and the structural funds from Charité-Universitätsmedizin Berlin allotted to Prof. Philipp Sterzer.

Research Articles

- Varrier, R. S., Stuke, H., Guggenmos, M., & Sterzer, P. (2019). Sustained effects of corrupted feedback on perceptual inference. *Scientific Reports*, *9*(1), 5537.

- Varrier, R. S., Rothkirch, M., Stuke, H., Guggenmos, M., & Sterzer, P. (*submitted*). Unreliable feedback deteriorates information processing in primary visual cortex.

To my parents

M S Varrier and K V Sreedevi,

for their continued support and encouragement.

Table of Contents

Abstract

To survive and adapt in an ever-changing world, we continuously sample our surroundings through our senses. To make correct inferences about the causes of sensory signals, it is essential to learn not only about the entities present around us, but also about the reliability of sensory information itself. The influential Bayesian brain hypothesis proposes that perception is an inferential process, depending not only on sensory data, but also on beliefs about the probable causes of sensory data. Over time, the brain arrives at estimates of what to expect and how much to rely on prior beliefs and sensory data in an environment. Feedback from the environment improves learning, thereby helping the brain to arrive at these estimates.

Now what will happen if the environmental feedback becomes unreliable? Providing unreliable feedback has previously been shown to impair task performance and increase pattern perception in noise. However, the mechanistic understanding about its underlying processes is currently limited. In this thesis, we explored the effects of unreliable feedback within the framework of Bayesian inference. We predicted that unreliable feedback would induce beliefs about the reliability of sensory information and lead the brain to down-weigh sensory data. To test this, we conducted two studies comprising visual orientation detection or discrimination tasks: Study I comprised two behavioural experiments and a simulation, and Study II comprised a neuroimaging experiment with functional magnetic resonance imaging (fMRI). We hypothesised that in both studies, the sensory data would be down-weighed after an unreliable feedback phase in which invalid feedback was given in half of the trials. As a result, we predicted that task performance would deteriorate. In Study I, we also predicted that in the presence of prior beliefs – induced by predictive cues on each trial – perceptual inference would shift towards the priors, leading to a higher number of cue-congruent responses. Further, simulations – which followed the experimental design of Study I – were performed to predict behaviour as a result of unreliable feedback. In one of the behavioural experiments of Study I, we additionally measured the confidence placed on responses in order to study changes in metacognitive awareness of performance. In Study II, we investigated whether the sensory data representations in the primary visual cortex (V1) deteriorate as a result of unreliable feedback. Reliable, correct feedback was used as a control condition in all the experiments.

Data from both studies demonstrated that performance did indeed decrease following unreliable feedback compared to reliable feedback. Moreover, observers increasingly relied on prior information as the feedback about their percepts became unreliable. At the neural level, fMRI showed that low-level stimulus representations deteriorated in V1 with unreliable feedback. To sum up, our results show that inducing beliefs about the reliability of sensory information by manipulating performance feedback can systematically influence perceptual inference and that these changes manifest at the earliest stages of cortical sensory processing.

Zusammenfassung

Um in volatilen und dynamischen Lebensumwelten zu überleben, erforschen wir unsere Umgebung kontinuierlich mithilfe unserer Sinnesorgane. Um verlässliche Rückschlüsse auf die Ursachen sensorischer Signale zu ziehen, ist es wichtig, nicht nur die uns umgebenden Entitäten zu lernen, sondern auch die Zuverlässigkeit der sensorischen Informationen selbst zu kennen. Laut der „Bayesian Brain"-Hypothese ist Wahrnehmung ein inferentieller Prozess, der nicht nur von sensorischen Daten abhängt, sondern auch von Vorannahmen über wahrscheinliche Ursachen sensorischer Daten. Im Laufe der Zeit lernt das Gehirn dabei einzuschätzen, inwieweit es sich auf Vorannahmen und sensorische Daten in einer Umgebung verlassen kann. Feedback aus der Umgebung verbessert das Lernen und hilft dem Gehirn, diese Schätzungen zu präzisieren.

Was passiert nun, wenn solches Feedback unzuverlässig ist? In vorherigen Arbeiten wurde gezeigt, dass unzuverlässiges Feedback die perzeptuelle Genauigkeit beeinträchtigt und Fehlwahrnehmungen in Rauschsignalen erhöht. Das mechanistische Verständnis über die zugrundeliegenden Prozesse ist jedoch derzeit sehr limitiert. Untersuchungsgegenstand der vorliegenden Dissertation ist es, die Auswirkungen von unzuverlässigem Feedback im Rahmen der Bayesischen Inferenz zu untersuchen. Unsere Hypothese war, dass unzuverlässiges Feedback unsere Einschätzung der Zuverlässigkeit sensorischer Informationen beeinflusst und dazu führt, dass sensorische Daten im Inferenzprozess weniger stark gewichtet werden. Hierzu wurden zwei Studien mit visuellen Reizen durchgeführt: Studie I umfasste zwei Verhaltensexperimente und eine Computersimulation; Studie II umfasste ein Experiment mit funktioneller Magnetresonanztomographie (fMRT). Unter der Annahme einer Abwertung sensorischer Information infolge unzuverlässigen Feedbacks wurde eine Verringerung der perzeptuellen Leistung vorhergesagt. In Studie I wurde zusätzlich prognostiziert, dass sich bei Vorliegen von Vorannahmen über wahrscheinliche Wahrnehmungsinhalte – experimentell induziert durch prädiktive Hinweise –, die Wahrnehmungsinferenz in Richtung dieser Vorannahmen verschieben würde, was sich in einer höheren Wahrscheinlichkeit hinweis-kongruenter Antworten niederschlagen würde. Beide Effekte wurden auch mithilfe einer Simulation analog zum experimentellen Design der Studie I untersucht. In einem der Verhaltensexperimente der Studie I wurde zusätzlich ein kontinuierliches Maß der Entscheidungssicherheit gemessen, um Veränderungen im

metakognitiven Bewusstsein nachzugehen. Auf neuronaler Ebene wurde schließlich in Studie II untersucht, ob sich sensorischen Datenrepräsentationen im primären visuellen Kortex (V1) als Folge unzuverlässigen Feedbacks verschlechtern würden. In allen Experimenten wurde in einer Kontrollbedingung zuverlässiges Feedback gegeben.

Die Ergebnisse beider Studien zeigen, dass die perzeptuelle Leistung nach unzuverlässigem Feedback im Vergleich zu einer Bedingung mit zuverlässigem Feedback abnimmt. Darüber hinaus verließen sich die Probanden zunehmend auf Vorannahmen über Wahrnehmungsinhalte. Auf neuronaler Ebene zeigte sich eine zunehmende Verrauschung sensorischer Repräsentationen in V1 als Folge von unzuverlässigem Feedback. Zusammenfassend zeigen die Ergebnisse, dass die Induzierung von Überzeugungen über die Zuverlässigkeit sensorischer Informationen durch manipuliertes Leistungsfeedback einen systematischen Einfluss auf perzeptuelle Inferenz hat und dass sich diese Veränderungen in frühen kortikalen Arealen sensorischer Verarbeitung manifestieren.

1. General Introduction

1.1. Background

We have evolved to live in a complex and dynamic environment, where adaptation is critical to survival. Here, the ability to identify salient information in the surroundings based on knowledge about the world is a potent adaptive tool for all species. We seldom see objects in their entirety but perceive them as wholes nevertheless. This is because we perceive our surroundings based on not only our sensory input, but also prior knowledge about their probable causes. This idea has been best formalised under the influential Bayesian inference theory. First proposed by Hermann von Helmholtz in 1867, this theory – then called "unconscious inference" – proposes that our percepts depend not only on sensory information, but also on prior knowledge and expectations. The balance between these two sources of information is vital, and has been explored in several studies not only of the healthy brain, but also to understand disorders including schizophrenia, anxiety, depression, visual neglect and autism (Karvelis, Seitz, Lawrie, & Seriès, 2018; Parr, Rees, & Friston, 2018; Paulus & Yu, 2012; Sterzer et al., 2018). The extent to which we rely on sensory inputs and prior knowledge depends critically on their reliability too. For instance, we trust our visual inputs more during the day than at night. This is not only because it is much harder to see at night, but also because our visual input depends more on subjective factors such as the light source (e.g.: moonlight, streetlight, light from incoming vehicles on the road etc.). We often use prior knowledge to navigate in such an environment; but in addition, critical teaching signals are provided by the consequences of our action or feedback – for instance, feedback is provided by the feeling under our feet as we walk (somatosensory feedback), how a target changes as we approach it (visual feedback) or even what an observer tells us about our position and gait (verbal feedback). Thus, feedback can be a useful tool to study the dynamics between prior beliefs and sensory information under conditions of uncertainty.

1.1.1. Feedback and learning

Feedback is ubiquitous in the real world and guides perception and subsequent action. For example, it is well-known that feedback is used to improve precision of our own actions (Fulvio & Rokers, 2017; Scott, 2004; van Vugt & Tillmann, 2015). As per the Law of Effect (Thorndike, 1913), the connections between situations and responses would get

reinforced when they were followed by "a satisfying state of affairs" and weakened if they were followed by "an annoying state of affairs". Thus, positive feedback reinforces the behaviour that leads to it and negative behaviour discourages it. Importantly, feedback – irrespective of its salience – reduces uncertainty, and its informational value is further highlighted by the increased brain activity in the caudate nucleus and the ventral striatum (Lempert & Tricomi, 2015), brain regions that have been previously associated with reward prediction errors (O'Doherty et al., 2004; Pagnoni, Zink, Montague, & Berns, 2002; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006).

Numerous studies have shown that feedback improves perceptual performance, especially when the task at hand is demanding (Herzog & Fahle, 1997; Liu, Lu, & Dosher, 2010, 2012). Perceptual learning has been proposed to be a type of reward-based learning, where although feedback is not necessary, it has informational value (Lempert & Tricomi, 2015) and can improve perception and learning (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Herzog, Aberg, Frémaux, Gerstner, & Sprekeler, 2012; Petrov, Dosher, & Lu, 2006; Seitz, Nanez, Holloway, Tsushima, & Watanabe, 2006). The brain areas that respond to cognitive feedback were also found to be largely the same as those involved in reward-based learning (Daniel & Pollmann, 2010). Feedback-related negativity (FRN), the electro-encephalographic (EEG) signature of negative performance feedback, encodes the feedback prediction error (Chase, Swainson, Durham, Benham, & Cools, 2010; Pfabigan, Alexopoulos, Bauer, & Sailer, 2011; Ullsperger, Fischer, Nigbur, & Endrass, 2014) and is sensitive to the salience (good vs. bad) of outcomes, not its magnitude (Hajcak, Moser, Holroyd, & Simons, 2006). Further, its amplitude has been shown to be proportionate to the probability of error correction (de Bruijn, Mars, & Hester, 2019) . Thus, existing studies indicate that cognitive feedback plays an important role in guiding learning.

1.1.2. Feedback validity and learning

Owing to its immense importance in perception and learning, the reliability or validity of feedback is critical. Previous studies have shown that feedback that was uncorrelated with performance could slow down learning or induce aberrant pattern perception in noise or noisy images. Different mechanisms have been proposed for this. One such phenomenon was that unreliable feedback induces a top-down "lack of control" (LOC),

where when individuals are unable to gain a sense of control objectively, they desire more structure, and this results in a greater propensity to see structure in noise (Vannucci, Mazzoni, & Cartocci, 2011; Whitson & Galinsky, 2008). The aforementioned studies did not observe a change in the overall error rate, but rather a higher rate of pattern identification in noisy images. Further, in these studies, independent tasks with different stimuli were used to induce LOC and to measure its effects on perception. Hence, a top-down intervention is likely to have influenced perception. Other studies have used two-choice identification tasks where participants had to identify either the offset direction of visual Vernier stimuli (Aberg & Herzog, 2012; Herzog & Fahle, 1997, 1999) or detect the presence of anomalous notes in melodies (Vuvan, Zendel, & Peretz, 2018). These studies showed that unreliable feedback decreases task performance across time or relative to reliable feedback, and that it could change the perceptual decision threshold. Contrary to these negative effects, a third set of studies showed that manipulating feedback could sometimes be beneficial. For instance, providing feedback with a positive bias could improve performance, since learning rates were to be boosted by the larger performance gradient given by the positively biased feedback (Shibata, Yamagishi, Ishii, & Kawato, 2009). In another study, providing fake feedback about stimuli that were not presented led to perceptual learning – indicating that feedback can facilitate learning purely based on top-down signals (Choi & Watanabe, 2012).

Thus, previous studies have manipulated feedback in different ways to study different aspects of perception and top-down learning. All of them are likely to have led to some kind of uncertainty, but *random* feedback (that has only a chance probability of being correct) simulates uncertainty about sensory data the most, since as we have seen above, any biased information was quickly learnt and was used to perform better in that situation. The studies that provided random feedback have shown either (1) an increased pattern perception without a change in overall performance or (2) a reversible deterioration in task performance. In this thesis, we attempt to provide a mechanistic understanding about how the perceptual inference adapts in a volatile environment by systematically providing unreliable, random feedback and by measuring its subsequent effects on behaviour and in the sensory cortex.

1.2. Theoretical framework

The goal of this thesis was to explain how perception shifts in a volatile environment as simulated by random, unreliable feedback. Such feedback would have no directional bias, discouraging potential response biases and learning strategies. In two-choice tasks, unreliable feedback would mean that valid feedback should be delivered in 50% of the trials and invalid feedback in the other 50% of trials. Thus, incorrect as well as correct choices (responses) would have equal probabilities of being followed by positive or negative feedback, which might make sensory data look less useful and could potentially lead to it getting down-weighed. In order to study this process, we mainly use two methodological approaches: (1) the Bayesian brain hypothesis to study changes on perceptual decisions especially in presence of prior beliefs, and (2) stimulus representations in sensory areas of the brain. Additionally, we studied changes in subjective awareness of performance as a result of unreliable feedback by means of confidence ratings about perceptual choices.

1.2.1. Bayesian inference and feedback

Perception has been proposed to be a Bayesian inferential process and hence depends both on the incoming sensory information and prior beliefs. The resultant posterior distribution determines the perceptual outcome (Hohwy, 2012; O'Reilly, Jbabdi, & Behrens, 2012).

This idea is named after Bayes' theorem on conditional probability, where the probability of an event is weighted by the prior knowledge about it. In an analogous manner, the incoming sensory signals are weighted by our prior knowledge about the probability of its occurrence. In Bayesian terms, the incoming sensory information is called the likelihood of the data given the hypothesis, whereas the prior is the probability of the hypothesis itself. The final percept corresponds to the probability of the hypothesis, given the data, and this is called the posterior probability. These three can be represented as probability distributions, based on the samples acquired over time. In the earlier example of navigating at night, in order to estimate the probability of an obstacle on the path, the likelihood would, for instance, be the probability of visually spotting an obstacle when there is one, i.e. $P(V|O)$, and the prior would be the probability of the occurrence of

the obstacle, i.e., P(O). The posterior would be the probability of the obstacle given the visual information or P(O|V). In Bayesian inference, P(O|V) is related to the prior and the likelihood as follows:

$$P(O|V) \propto P(V|O) * P(O) \qquad (1.1)$$

The influence of likelihood and prior on the posterior, which in turn influences both the perceptual decisions and updates the prior, depends on two critical characteristics of the two probability distributions – namely their means (μ) and variances (v), also known as their first and second moments, respectively. In particular, the lower the variance of a distribution, the higher the precision (1/v) of that distribution – and consequently, higher is its contribution to the posterior. For instance, if the sensory data varied very little over time, the corresponding likelihood probability distribution would have low variance and would consequently have a larger influence on the posterior.

In the real world, the brain continuously learns from the surroundings, thus updating the priors about sensory information as well as how much to weigh sensory data and prior beliefs. These weights in turn correspond to the precision, or the inverse of the variance, of the respective probability distributions (Adams, Stephan, Brown, Frith, & Friston, 2013; Knill & Pouget, 2004). Feedback helps us to arrive at the precise estimates of the moments of the distributions (means and variances) corresponding to each source of information. When feedback is manipulated by providing random feedback in half of the trials, data points are misclassified in half of the trials, i.e., several wrong data points are included in each stimulus distribution. As a result of this, the perception and consequently decisions about stimuli would get noisier, thus not only slowing, but also potentially impairing learning. Under such conditions, priors could potentially become relatively more salient.

These questions were probed in **Study I**. As part of **Study I**, two behavioural experiments consisting of visual orientation detection and discrimination tasks coupled with probabilistic priors were performed. In addition, a simulation was performed to better explain how a Bayesian observer would make inferences when provided with unreliable feedback.

1.2.2. Sensory representation in the brain

Theories of precision-based prediction error (Adams et al., 2013) propose that the prediction error, i.e., the difference between predicted and actual sensory data, is weighted by its precision before being used to update the knowledge about stimuli, or prior. This is analogous to the precision of the likelihood distribution in classical Bayesian inference, since in both cases, the posterior is shifted in proportion to the precision of the sensory evidence or the residual signal that is not "explained away" by the top-down expectation (den Ouden, Daunizeau, Roiser, Friston, & Stephan, 2010).

We predicted that unreliable feedback would lead the brain to down-weigh sensory data in a top-down manner, and that as a result of this, the representational accuracy of stimulus information in the sensory areas would decrease. Previous studies have shown that higher sensory uncertainties induced by directly varying the signal contrast or signal-to-noise ratio can deteriorate multivariate representations of stimuli in the sensory areas of the brain (Darcy, Sterzer, & Hesselmann, 2019; Hebart, Donner, & Haynes, 2012; Ludwig, Sterzer, Kathmann, & Hesselmann, 2016; Tong, Harrison, Dewey, & Kamitani, 2012). However, it has been shown that prior beliefs about stimuli can also influence activity and stimulus representation in the early and intermediate-level visual areas (Den Ouden, Friston, Daw, McIntosh, & Stephan, 2008; Jiang, Summerfield, & Egner, 2013; Kok, Brouwer, Gerven, & Lange, 2013; Kok, Jehee, & de Lange, 2012; Schmack et al., 2013). Yet another study has shown that likelihood uncertainty activated sensory areas whereas prior uncertainty activated higher brain areas (Vilares, Howard, Fernandes, Gottfried, & Kording, 2012). Further, attention has been long known to enhance brain activity in sensory areas (Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999; Kastner, Weerd, Desimone, & Ungerleider, 1998; Moran & Desimone, 1985; Roelfsema, Lamme, & Spekreijse, 1998; Treue & Trujillo, 1999). Thus, activity and representations in the sensory cortex are influenced both by sensory uncertainty and the top-down processes.

Based on these observations, we designed **Study II** to investigate the effects of unreliable feedback on stimulus representations in the sensory cortex, by means of an orientation discrimination task and functional magnetic resonance imaging (fMRI).

1.2.3. Metacognitive awareness

Metacognitive awareness about perceptual performance refers to the awareness about one's own performance. This is typically measured using confidence ratings on each trial (Kleitman & Stankov, 2007; Stankov, 2000; Yeung & Summerfield, 2012). Confidence ratings are known to increase in proportion to the task performance accuracy as well as the differences between the sensory stimuli being compared (Daniel & Pollmann, 2012; Guggenmos, Wilbertz, Hebart, & Sterzer, 2016; Peirce & Jastrow, 1884; Vickers, 2014).

As discussed in Section 1.1.1. (p.6-7), external feedback guides learning, although it is not necessary. Recently, a few studies have shown that in the absence of external feedback, confidence additionally functions as reinforcement signals and guides learning (Daniel & Pollmann, 2012; Guggenmos et al., 2016; Hebart, Schriever, Donner, & Haynes, 2016). On the flip side, imprecise feedback has been shown to affect the precision of confidence – for instance, participants have been shown to be overconfident about their own relative to others' performance when noisy (but unbiased) feedback was provided (Grossman & Owens, 2010).

Since metacognitive awareness of performance has been shown to decrease along with task difficulty as well as uncertainty, the delivery of unreliable feedback would likely decrease confidence, as was shown in a previous study (Vuvan et al., 2018). An alternate possibility is that confidence would not decrease in line with the predicted decrease in precision of sensory data (Section 1.2.1., p.10). This would result in a relative overconfidence that would then accompany the decisional shift towards prior beliefs. Overconfidence about prior beliefs is typical in perceptual disorders with impaired bottom-up sensory processing such as schizophrenia (Balzan, 2016; Köther et al., 2012; Moritz et al., 2014).

To understand the effects of unreliable feedback on subjective awareness, confidence ratings were collected in one of the two behavioural experiments of **Study I**.

1.3. Approach

As discussed in Section 1.1.2. (p.7–8), previous studies have shown that manipulating feedback validity could increase pattern perception, impair learning and confidence, shift decision criteria or even improve performance. However, the underlying mechanism could have varied depending on stimuli, task demands and the way in which feedback was manipulated.

The goal of this thesis is to induce an uncertain state by delivering unreliable performance feedback in perceptual tasks and to study its effect on perceptual inference using behavioural and neuroimaging techniques. To study the effects of unreliable feedback, we use two-choice visual detection or discrimination tasks, and provide feedback that is valid at chance-level, i.e., 50%. As a control condition, we used reliable feedback which was 100% valid. To avoid carryover effects between the two types of feedback (reliable/ unreliable) in these experiments, the control session was performed during a separate session (day) either by the same participants (Study I) or by a different groups of participants (Study II).

In both studies, unreliable or reliable feedback was provided in dedicated "intervention" phases. The sustained effects of feedback manipulation were measured in "test" phases, which preceded and followed each intervention phase. A key aspect to ensuring successful feedback manipulation was for it (the manipulation) to not be too obvious to participants. To this end, the valid and the invalid feedback trials were interspersed randomly in the intervention phases with unreliable feedback. In addition, the perceptual tasks were moderately difficult and the stimuli were set to a task difficulty of 80% at the beginning of an experimental session. The performance threshold was not set to a higher difficulty level to avoid a floor effect. In Study I, the difficulty was adjusted by varying the signal-to-noise ratio of visual gratings in detection (experiment 1) and discrimination (experiment 2) tasks, and priors were induced using probabilistic cues. In Study II, a different orientation discrimination task was used while brain images were acquired using fMRI. The stimuli consisted of high-contrast visual gratings, and task difficulty was adjusted by varying the deviation of the gratings from the two diagonal reference orientations. In order to measure the changes in metacognitive awareness, confidence ratings were collected in addition to perceptual choices in experiment 2 of Study I. In the remaining experiments, only the binary responses were collected. Lastly,

the behavioural effects of unreliable feedback observed in the two experiments of Study I were replicated using simulations.

The effect of unreliable feedback relative to reliable feedback was measured by comparing task performances (Studies I and II), the cue-congruence of responses (Study I), the stimulus representations in primary visual cortex (Study II) and the metacognitive awareness of performance (Study I). In both the studies, the primary goal was to study how these variables changed between pre- and post-intervention phases in the unreliable feedback sessions compared to the reliable feedback sessions.

1.4. Hypotheses

We predicted that unreliable feedback would exacerbate the uncertainty about sensory data, since feedback was often incongruent to the responses. In Bayesian terms, unreliable feedback decreases the precision of stimulus distributions, leading the brain to down-weigh sensory information and up-weigh prior information, and consequently shifting the posterior distribution – and consequently perception – towards prior beliefs. Behaviourally, this predicts a higher number of errors in perception and a higher number of prior-congruent responses. Further, we predicted that down-weighing sensory information would decrease the precision of sensory representations in early sensory areas. Lastly, we predicted that the self-rated confidence, a measure of subjective awareness of performance, would also decrease in line with objective performance.

The primary hypotheses are formulated more specifically below:

(1) Unreliable feedback leads to a decrease in task performance accuracy (Study I, experiments 1,2 and simulations; Study II)

(2) Unreliable feedback shifts perceptual inference away from sensory data and towards learned priors (Study I, experiments 1,2 and simulations)

(3) Unreliable feedback deteriorates stimulus representations at early stages of visual processing in the cortex, i.e., V1 (Study II)

In addition, there was an additional, exploratory hypothesis, which is given below:

(4) Unreliable feedback decreases the metacognitive awareness of one's own performance (Study I, experiment 2)

Studies I and **II** are described in detail in Chapters 2 (p.16) and 3 (p.49), respectively.

2. Study I: Sustained Effects of Unreliable Feedback on Perceptual Inference

The work presented here has been published as:

- Varrier, R. S., Stuke, H., Guggenmos, M., & Sterzer, P. (2019). Sustained effects of corrupted feedback on perceptual inference. *Scientific Reports*, *9*(1), 5537.

2.1. Introduction

According to the Bayesian brain hypothesis, perception is an inferential process, and is the result of the integration of probability distributions representing our prior beliefs ("prior") and new sensory evidence ("likelihood") (Hohwy, 2012; O'Reilly et al., 2012). The balance between the prior and the likelihood is thought to be dynamically adjusted as we continuously update our estimates about their reliability (Adams et al., 2013; Knill & Pouget, 2004). Feedback aids learning by providing the outcomes of responses, especially when the task is challenging (Herzog & Fahle, 1997; Liu et al., 2010, 2012). In this study, we sought to investigate the effects of the reliability of feedback on Bayesian perceptual inference.

Previous studies have already shown that perception was negatively affected when unreliable feedback, i.e., feedback that is not correlated with performance, was delivered in visual tasks. Relative to reliable feedback, task performance often decreased in two-choice decision-making tasks with visual or auditory stimuli (Aberg & Herzog, 2012; Herzog & Fahle, 1997, 1999; Vuvan et al., 2018). Further, unreliable feedback also led to higher pattern perception in object-naming tasks that used noisy images (Vannucci et al., 2011; Whitson & Galinsky, 2008). These findings can be understood within the perceptual inference framework. For instance, repeatedly misclassifying stimuli would reduce the precision of the likelihood distributions of sensory data, as a result of which the task performance would deteriorate. This in turn would make the brain up-weigh available prior information, making them relatively more precise. Consequently, perceptual inferences shift towards priors, leading to heightened pattern perception in noise when asked to identify objects. Since these studies did not have clearly defined priors, participants are likely to have used priors about objects learnt over the course of life.

Our goal in this study was to investigate how unreliable feedback would influence perceptual inference when probabilistic beliefs about the stimuli are induced during the experimental session. We contested that when unreliable feedback (sometimes correct, sometimes incorrect) gets delivered repeatedly to the same sensory data, its representations become corrupted or less precise. Consequently, perceptual inference would shift towards available priors. The compensatory role of priors as a result of imprecise feedback has already been shown in a previous study, where increasing

feedback uncertainty in a visuo-motor task led to a stronger influence of prior beliefs on behaviour (Körding & Wolpert, 2004).

Metacognitive awareness of performance, typically measured using confidence ratings, is a well-known indicator of one's awareness of their own performance. A previous study showed that unreliable feedback could impair self-rated confidence (Vuvan et al., 2018). In this study, as a secondary research question, we also investigated the changes in confidence as a result of unreliable feedback.

This chapter describes two behavioural experiments and a simulation. A key aspect of our investigation was to deliver unreliable feedback on perceptual performance when only the sensory evidence (and no additional predictive information) was available, and to then measure its effects in *subsequent* runs where two sources of information were provided on each trial – (1) a learned predictive cue inducing a prior belief and (2) the actual sensory evidence. The experiments used challenging visual detection (experiment 1) and discrimination (experiment 2) tasks. Each participant took part in two experimental sessions (Figure 2.1, p.20). In the "Intervention" runs of each session, participants received either unreliable feedback (valid in one half of the trials and faulty in the other half) or reliable feedback (valid in all of the trials). The session with reliable feedback interventions served as the control session. The sustained effects of these interventions were measured in "Test runs" that were interleaved between the intervention runs. In the test runs, probabilistic cues were presented before each stimulus. The test runs received reliable feedback on both the sessions.

Lastly, computer-generated simulations of unreliable feedback were implemented, and their effects on perception was measured thereafter. The simulations followed the design of the behavioural experiments and consisted of intervention and test "runs" with unreliable/reliable feedback and predictive cues, respectively.

2.2. Hypothesis

We reasoned that unreliable feedback on perceptual performance would lead to erroneous updating of likelihood distributions, rendering them more imprecise over time. Therefore, our hypotheses were that, as a result (1) task performance would decrease, and (2) when predictive information becomes available, there would be an increased

reliance on this information. Further, in experiment 2, we had the exploratory hypothesis that confidence would decrease as a result of unreliable feedback, in parallel with performance.

2.3. Behavioural experiments

2.3.1. Materials and methods

As both experiments were very similar in experimental design, they are described together here, and distinctions are made wherever the methodology differed.

2.3.1.1. General study design

In both the experiments, each participant took part in two sessions on separate days – one session consisted of intervention runs with unreliable feedback, and the other correspondingly had reliable feedback (Figure 2.1a, p.20). Thus, the only difference between the two sessions was the presence of unreliable feedback in the intervention runs of one session and reliable feedback in the corresponding intervention runs of the other session. The order of sessions was counter-balanced across participants.

Each session started with preliminary *training* runs to facilitate learning of the priors, which was then followed by the *threshold estimation* runs in order to set the perceptual threshold for the *main experiment*. This was then followed by the main experiment, comprising intervention runs to deliver unreliable (or reliable) feedback regarding the perceptual choice, and test runs to measure the sustained effects of the unreliable (or reliable) feedback interventions (Figure 2.1b-c).

2.3.1.2. Timeline of an experimental session

Each session lasted for about two hours, including the time taken for the breaks, task instructions, training, threshold estimation and debriefing besides the main experiment. The main experiment (four test runs and three intervention runs, Figure 2.1c) lasted for approximately 70 minutes, with each test run lasting for approximately 6 minutes and

each intervention run lasting for approximately 9 minutes. Participants were encouraged to take short breaks between runs in order to minimise the effects of fatigue on behaviour.

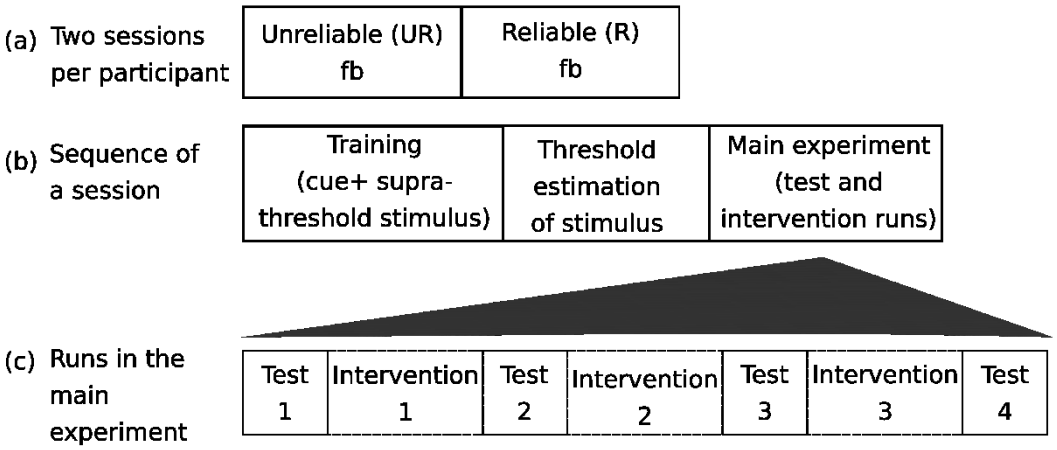

Figure 2.1. (a) Each participant took part in two sessions, one with unreliable and one with reliable feedback. (b) Each session consisted of three parts: training, threshold estimation and the main experiment. (c) The main experiment comprised four test runs interleaved with three intervention runs. The intervention runs were used to deliver the unreliable (or reliable) feedback, and the test runs measured its lasting effects.

2.3.1.3. Participants

The study was approved by the ethics committee at Charité - Universitätsmedizin Berlin, and informed consents were collected from all participants. All the methods were carried out in accordance with the relevant guidelines and regulations. 37 participants (six male, ages 27.7±5.9) took part in experiment 1, three of whom had to be excluded due to technical difficulties (i.e. final N = 34). A different group of 32 participants (seven male, ages 24.75±3.6) took part in experiment 2. In both the experiments, every participant took part in two sessions corresponding to unreliable and reliable feedback. The mean time between the two sessions were 1.94 days (*SD* = 1.98 days) in experiment 1 and 1.7 days (*SD* = 1.7 days) in experiment 2.

2.3.1.4. Stimuli

Images were constructed from an overlay of annular gratings and noise images (Figure 2.2a, p.22). Annular gratings at an orientation of 45° counterclockwise or clockwise (the latter only for experiment 2) were generated such that the spatial frequency of the gratings would be 0.87 cycles/degree, the inner diameter of stimuli 9.94° and the outer diameter 20.93°. Noise images were generated by performing spatial smoothing of a two-dimensional annular noisy patch of the same inner and outer diameters as that of the gratings. Next, based on a previous study with noisy gratings (Guggenmos et al., 2016), the grating and noise images were combined in the following manner for all the test and intervention runs of the main experiment (Figure 2.1c, p.20) :

$$\boldsymbol{I} = 0.5\ (1 + w_s \cdot \boldsymbol{G} + w_n \cdot \boldsymbol{N}) \qquad (2.1)$$

where $\boldsymbol{G}$ and $\boldsymbol{N}$ were two-dimensional matrices consisting of the grating and smoothed noise images respectively, scaled to the interval [−0.5, 0.5], and $\boldsymbol{I}$ was the resultant image matrix. Parameters w_s and w_n were signal and noise weights respectively. The parameter w_n was maintained at a constant value of 0.25 across subjects and sessions, and w_s was set based on the signal threshold s (in percent) estimated prior to the main task for each participant during each session as follows:

$$w_s = w_n \cdot \frac{s}{100 - s}. \qquad (2.2)$$

2.3.1.5. Cues

Auditory pure tones of high and low frequencies (1000Hz and 300Hz respectively) adjusted for loudness were used as cues, in line with previous studies that used audio-visual associative learning cues to study the influence of priors on behaviour (Iglesias et al., 2013; Kok et al., 2012; Schmack, Weilnhammer, Heinzle, Stephan, & Sterzer, 2016) . On each trial, a cue tone was played for 300ms, and after a brief interval (1000ms in experiment 1, 500ms in experiment 2), the visual stimulus was presented. The tones were probabilistically coupled to stimuli of one type in 75% of the trials and with stimuli of the other type in 25% of the trials. The type of cue-stimulus association (type 1: HighTone/Stim1 and LowTone/Stim2; type 2: HighTone/Stim2 and LowTone/Stim1) remained constant for each participant across sessions, and this was balanced between

participants. Participants were instructed to pay attention to the tones in the runs that had them; they were told that these could be helpful. Participants were not informed as to *how* useful the cue would be or whether the cue-stimulus association would change over time. This was to be learnt by them over the course of the experiment.

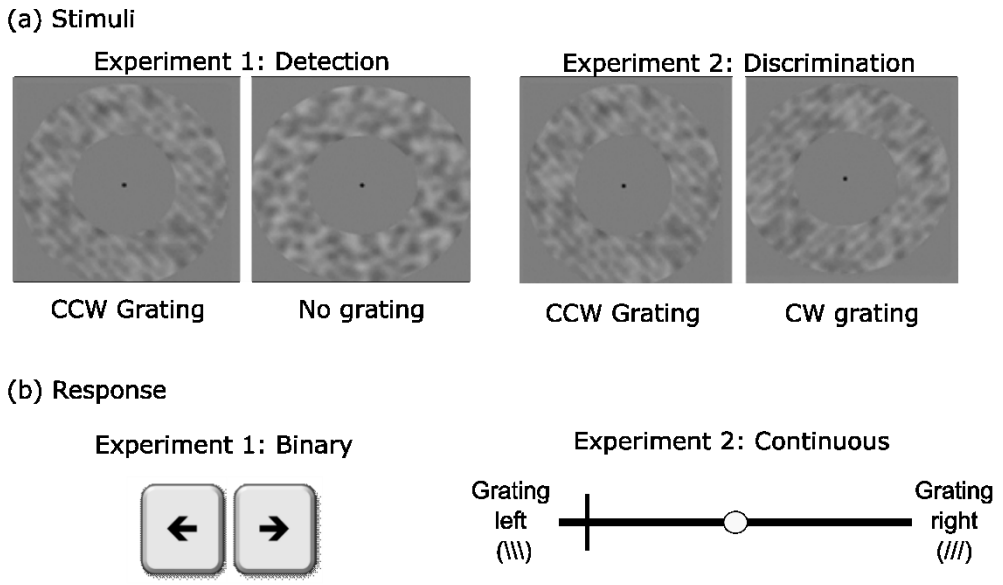

Figure 2.2: (a) Visual stimuli (CCW: counter-clockwise, CW: clockwise), (b) response options, (c) visual feedback, and (d-e) trials in cued and un-cued runs (i.e., test and intervention runs, respectively) for the behavioural experiments. R= reliable feedback (valid in 100% of trials), UR = unreliable feedback (valid in 50% of trials).

2.3.1.6. Feedback

Trial-by-trial visual feedback was delivered at the centre of the screen, in line with previous studies that have used colour-coded or symbolic cues (Been, Jans, & De Weerd, 2011; Lempert & Tricomi, 2015; Miltner, Braun, & Coles, 1997). An upward-pointing green triangle indicated a correct response and a downward-pointing red triangle indicated an incorrect response (equilateral triangles with 0.78° edges, see Figure 2.2c, p.22). In runs with unreliable feedback (i.e., intervention runs in the session with unreliable feedback), the presentation of the red/green triangles was pseudo-randomised, such that in half of the trials of each stimulus type, the feedback delivered was faulty.

2.3.1.7. Experimental procedure

The task was implemented using PsychToolbox 3.0.11 (psychtoolbox.org) on a computer screen (resolution: 1280x960 pixels, refresh rate: 60 Hz) placed 46 cm away from the chinrest, where the participant was positioned. Participants were instructed to fixate at the centre of the screen throughout the experiment, where a black dot (radius: 0.34° visual angle) was presented at all times *except* during feedback delivery, when the feedback (Figure 2.2c, p.22) replaced the dot. In experiment 1, the participants' task was to report the presence or absence of gratings using the left and right arrow keys on a standard German keyboard. In experiment 2, the task was to report both the perceived orientation of gratings and the confidence about the response on a linear visual analogue scale using a single mouse-click (Figure 2.2b right, p.22). The left and right halves of this scale corresponded to the perception of counter-clockwise and clockwise perception of gratings, respectively. The distance from the centre (white circle) indicated confidence, i.e., responses that were closer to the left and right tips of the scale indicated high levels of confidence about the respective percept, and those near the centre indicated low levels of confidence. It was not possible to click at the centre of the response bar, forcing participants to indicate a decision about the orientation to proceed. To minimise the effects of fatigue or laziness on confidence ratings, the initial position of the cursor was random on each trial. To reduce reporting errors in confidence, time restriction was not imposed in experiment 2.

2.3.1.8. Time course of a session

Each session consisted of three parts: training, threshold estimation and the main experiment (Figure 2.1b, p.20). These parts are described below.

(i) ***Training:*** In this part, an association was induced between auditory cues and visual stimuli. To facilitate this associative learning, supra-threshold stimuli (12% signal) were presented, and reliable feedback was given. There were three runs in the training phase, and each run consisted of 48 trials. The time-course of a trial in the training runs was similar to those of the test runs in the main experiment (Figure 2.2d, p.22).

(ii) ***Threshold estimation:*** A staircase procedure was used to determine the percentage of signal (grating) required to attain a performance level of 80% correct responses. To estimate the signal threshold *s* (see equation 2.2, p.21), a 2-down-1-up staircase procedure with two phases was performed before each session with a step-size down/step-size ratio of 0.5548 (García-Pérez, 1998; Guggenmos et al., 2016). The first phase was to determine the approximate signal threshold and had larger step-sizes (1% signal up, 0.5548% signal down). The second phase started at the threshold estimated by the first staircase and had smaller step-sizes (0.5% signal up, 0.2774% signal down). No auditory cues were presented during the staircase, but reliable feedback was provided. In experiment 1, the first and the second phases of the staircase proceeded until a certain number of reversals were attained (8 and 10 for phases 1 and 2, respectively), or 80 trials were completed. The signal threshold was determined based on the signal levels at which the last 4 reversals occurred in the second phase. Experiment 1 showed that six reversals were sufficient to arrive at the threshold signal. We therefore confined threshold estimation to six reversals in experiment 2, while keeping everything else the same as in experiment 1.

(iii) ***Main experiment:*** The main experiment in both the sessions comprised seven runs in total (Figure 2.1c, p.20). In all of them, visual stimuli were presented at the 80% performance threshold determined in the previous step. The test runs served to probe the sustained behavioural changes resulting from feedback manipulation. Each test run consisted of 64 trials. The trials here were similar to the training runs,

comprising predictive cues and reliable feedback (Figure 2.2d, p.22). In the intervention runs, either unreliable (50% valid) or reliable feedback (100% valid) was delivered. Feedback reliability was the same across the three intervention runs within a session. An intervention run consisted of 128 trials (Figure 2.2e, p.22). In experiment 2, at the end of each run, participants were asked to rate their motivation on a scale from 0 to 100.

2.3.1.9. Debriefing

At the end of the second session in both the experiments, participants were asked to fill a short questionnaire, which consisted of questions about their awareness of having received unreliable feedback. The relevant questions are written below (verbatim). Questions (iii) and (iv) were added to the questionnaire after experiment 1 (i.e., only in experiment 2), to get a more quantitative estimate of feedback manipulation awareness.

(i) How did you find the feedback (helpful/ confusing/ no difference etc.)?

- *a. On Day 1:*
- *b. On Day 2:*

(ii) Did you notice anything odd about the feedback? If so, on which session/ day did you notice it?

(iii) How reliable (correct) was the feedback? (0=absolutely not reliable (correct), 100=totally reliable (correct))

- *a. On day 1: 0% ----------100%*
- *b. On day 2: 0% ---------- 100%*

(iv) During the experiment, did you think that the feedback was manipulated? Please circle your response

- *a. On day 1: Definitely yes/ highly likely / maybe / highly unlikely/ definitely not*
- *b. On day 2: Definitely yes/ highly likely / maybe / highly unlikely/ definitely not*

2.3.1.10. Motivation ratings

After each run in the main experiment (Figure 2.1c, p.20) in experiment 2, participants were asked to rate their motivation on a scale from 0 to 100 (0% = not motivated at all, 100% = fully motivated).

2.3.1.11. Eye-tracking

To ensure fixation, a video-based eye-tracker (Cambridge Research Systems, UK; sampling rate: 250 Hz; spatial accuracy: 0.05°) was used throughout the experiment. A region of interest with radius 15mm (1.87°) was defined around the fixation. If the detected gaze position was outside this region, the trial would not start, and as a cautionary note to the participant, the fixation dot (●) would switch to a ring (○) of the same radius until gaze was returned to the fixation area. After stable fixation for 700ms, another 300ms interval followed, after which the auditory cue or visual stimulus was presented, depending on the run type. Fixation was monitored during the presentation of the visual stimuli as well in the test runs. In case the fixation was broken, (1) stimuli disappeared and (2) the fixation dot was replaced by a ring (○) at the centre, like the fixation check at the onset of each trial.

2.3.2. Data Analysis

2.3.2.1. Responses

The collected responses were binary in experiment 1 and continuous in experiment 2 (Figure 2.2b, p.22). Thus, responses in experiment 1 were stored as integer values 1 and 2 corresponding to the two percepts. On each trial in experiment 2, the response was stored as a decimal value within the interval [–1, 1], where the sign (– /+) indicated whether the grating was perceived to be tilted counter-clockwise or clockwise, and the absolute value indicated confidence, with higher values indicating greater confidence. It was not possible to click near the centre of the response bar in the interval [–0.03, 0.03]. Thus, participants had to make a perceptual choice on each trial.

2.3.2.2. Dependent variables

In order to test the main hypotheses, two dependent variables were computed for each test run: (1) performance accuracy, measured by the percentage of correct responses and (2) cue-congruent behaviour, measured by the cue-congruence index (CCI) which is defined below. Due to the correlation between stimuli and cues in the test runs (co-occurrence in 75% of the trials), a decrease in performance would be paralleled by a decrease in cue-congruence. To get a measure of cue-congruence that does not depend on this performance-related change, we computed the percentage of correct responses (CR) separately for cue-congruent (CC) and cue-incongruent (CI) trials within a run, and then defined CCI as a difference between them:

$$CCI = CR_{CC} - CR_{CI} \qquad (2.3)$$

Thus, cue-congruence, or CCI, increases if the performance in CC trials increases *relative* to that of CI trials. The upper bound for CCI is 100, and we would get this if cue-congruent responses were made on each trial. Similarly, a CCI value of 0 would indicate that performance was the same in congruent and incongruent trials – conveying that the cue had no influence on responses.

Next, in order to understand the overall changes in performance and cue-congruence across the test runs within a session, we fitted linear functions across time for each dependent variable and session, resulting in two slopes each for performance and cue-congruence per participant.

To study changes in subjective ratings of confidence in experiment 2, mean confidence was computed for each run, and then slopes of confidence were computed for each session and for each participant.

Since responses in both the behavioural experiments were un-speeded (i.e., participants were instructed to be as accurate as possible) and since experiment 2 used a continuous response scale in which the location of a bar had to be accurately adjusted (to obtain confidence ratings), we focused on response accuracy and did not analyse reaction times.

2.3.2.3. Statistical analysis

Our hypotheses were tested by means of two-way repeated measures analyses of variance (RM-ANOVAs) of the run-wise estimates of each dependent variable (i.e., performance and CCI). There were two within-subject factors, namely, feedback type (*fbtype* – unreliable/reliable) and time (*test run number* - 4 levels). The changes across time within the two sessions were compared using the linear interaction analysis measured by means of ANOVA contrasts with linear weights. The sequence of sessions (a binary value indicating whether unreliable feedback was delivered in the first or second session) and the cue-stimulus association type (types 1 or 2, indicating different combinations of cues and stimuli, see Section 2.3.1.5., p.21 for details) were included as between-subject factors, and the duration between sessions (number of days) was included as a between-subject covariate. One-sample t-tests of the session-wise slopes were also performed separately for the unreliable and reliable feedback sessions to understand the nature of changes across time (positive and negative slopes to indicate linear increases and decreases across time within a session, respectively). In experiment 2, additionally, the changes in mean confidence were studied using the same analyses. Lastly, to examine the common mechanisms underlying the two experiments (which were similar in design, but differed in the stimuli and response types, see Figure 2.2a-b on p.22), a post-hoc RM-ANOVA was performed on the pooled dataset with *experiment number* (i.e., 1 or 2) as an additional between-subject factor. IBM SPSS Statistics 23 and MATLAB R2013b were used for all statistical analyses.

2.3.2.4. Post-hoc analyses

We performed additional post-hoc analyses to test for relationships between the main effects of interest (decrease in performance and increase in CCI) and two subjective factors: one, the awareness of feedback manipulation and two, motivation scores (only experiment 2). To estimate the effects of interest, changes in performance and CCI were quantified as slope differences ($\delta_{\text{Perf. slope}}$ and $\delta_{\text{CCI slope}}$), defined as the differences between slopes for the unreliable and reliable feedback sessions for each variable. A third post-hoc test was performed to verify that the initial cursor position (which was randomly assigned on each trial) did not influence responses. Each of these analyses are described below.

(i) *Awareness of feedback manipulation:* Based on the answers to the debriefing questions, the awareness of feedback manipulation was encoded for each participant as either 0 (completely unaware, always trusted feedback), 0.5 (partially aware, noticed some oddity in feedback) or 1 (completely aware, realized that the delivered feedback was sometimes faulty). In both experiments, Spearman's rank correlation coefficients were computed between the awareness of feedback manipulation and the differences (unreliable – reliable) between session-wise slopes of performance ($\delta_{\text{Perf. slope}}$) and CCI ($\delta_{\text{CCI slope}}$) separately.

(ii) *Motivation ratings (experiment 2 only):* The percentage ratings of motivation in the test runs were fitted with linear functions separately for each test session, and slopes were compared using a paired t-test. Next, a slope difference ($\delta_{\text{Motiv. slope}}$) was computed between the unreliable and reliable feedback sessions. Correlation analyses were performed between this slope difference and analogous slope differences in performance and CCI ($\delta_{\text{Perf. slope}}$ and $\delta_{\text{CCI slope}}$) using Pearson's correlation.

(iii) *Initial position of the cursor (experiment 2 only):* To test whether the initial position of the cursor in the response bar (Figure 2.2b, p.22) could have influenced participants' responses, two tests were performed.

a. Overall correlation: The overall correlation between the initial cursor positions and participants' responses (i.e., their final cursor positions) was computed for each participant after pooling the data across sessions and test runs. This was then Z-transformed and tested for significance using a one-sampled t-test.

b. Correlation across time: To test whether correlations could have emerged during the experiment, correlations between the initial cursor positions and the responses or final cursor positions were computed separately for each test run and session, and then Z-transformed. The resultant Z-scores were then analysed using a two-way RM-ANOVA with the same within-subject and between-subject factors as were used to test the main hypotheses.

2.3.3. Results

The present study tested two main hypotheses and one secondary hypothesis about the sustained effects of unreliable feedback on perceptual decisions in a visual task (delivered in dedicated intervention runs). The main hypotheses were that in the test runs following the delivery of unreliable feedback, which also consisted of probabilistic priors, (1) task performance would deteriorate, and (2) responses would shift towards prior beliefs. The secondary hypothesis was that the metacognitive awareness, measured as self-rated confidence in responses, would deteriorate along with the impairment in perceptual inference.

2.3.3.1. Unreliable feedback impairs task performance

The critical analysis to assess the sustained effects of unreliable feedback on performance was the interaction between the within-subject factors fbtype (unreliable/reliable) and time (test runs 1 to 4). This was tested by means of two-way RM-ANOVAs.

Using a linear ANOVA contrast we tested the hypothesis that the linear performance change across time was different between the unreliable and reliable feedback conditions (henceforth referred to as *linear interaction effect*). In line with our hypothesis, we found a significant linear interaction effect between the factors feedback type (fbtype) and test run number (time) in experiment 1 ($F(1, 29) = 7.93$, $p = .01$; Figure 2.3a, p.32), but not in experiment 2 ($F(1, 27) = 1.5$, $p = .23$; Figure 2.3b, p.32). However, the session-wise slopes for performance were negative (i.e., decreased over time) in the *unreliable* feedback sessions of both the experiments (experiment 1: $M = -3.05$, $SE = 0.66$, $t(33) = -4.64$, $p < .001$; experiment 2: $M = -1.67$, $SE = 0.74$, $t(31) = -2.24$, $p = .03$), but the slope for the *reliable* feedback sessions did not deviate significantly from zero (experiment 1: $M = -0.05$, $SE = 0.54$, $t(33) = -0.09$, $p = .93$; experiment 2: $M = -0.33$, $SE = 0.71$, $t(31) = -0.46$, $p = .65$). The between subject factors (tone-stimulus association type, sequence of sessions) and covariate (number of days between sessions) that were included in the ANOVA tests did not show significant interactions with fbtype and time (all $p > .27$).

The pooled RM-ANOVA analysis performed to examine the common mechanisms underlying the two experiments (using experiment number as an additional between-

subject factor) revealed a significant linear interaction effect between time and fbtype ($F(1, 57) = 8.63$, $p = .005$; Figure 2.3c, p.32), and like in the individual experiments, resulted from a decline in performance in the pooled dataset in unreliable feedback session (slope $M = -2.38$, $SE = 0.5$, $t(65) = -4.78$, $p < .001$), but not in the reliable feedback session (slope $M = -0.18$, $SE = 0.44$, $t(65) = -0.42$, $p = .68$). The three-way linear interaction between fbtype, time and experiment number was not significant ($F(1, 57) = 2.4$, $p = .13$), indicating that the fbtype-by-time interaction was comparable across experiments.

Thus, unreliable feedback interventions led to a decrease in task performance in the ensuing test runs even though reliable feedback was provided in these runs.

Complementary analysis

In order to understand if a similar drop in performance could be seen *during* the delivery of unreliable feedback as well, changes in the intervention runs were studied using linear ANOVA contrasts with fbtype and time (intervention runs 1 to 3) as the within-subject factors of interest, and the sequence of sessions and the number of days between sessions included as the between-subject factor and covariate, respectively. Similar to the test runs, the intervention runs too showed significant linear interaction between the factors fbtype and time (experiment 1: $F(1, 31) = 7.67$, $p = .01$; experiment 2: $F(1,29) = 4.82$, $p = .04$, Figure 2.4a-b, solid circles and triangles, p.32), which was associated with a selective performance drop over time in the unreliable feedback session as evidenced by the significant negative slopes in these sessions (experiment 1: $M = -1.81$, $SE = 0.79$, $t(33) = -2.29$, $p = .03$; experiment 2: $M = -1.62$, $SE = 0.7$, $t(31) = -2.33$, $p = .03$), but not in the reliable feedback sessions (experiment 1: $M = 0.77$, $SE = 0.56$, $t(33) = 1.37$, $p = .18$; experiment 2: $M = 1.37$, $SE = 0.8$, $t(31) = 1.72$, $p = .1$). In both the experiments, three-way interactions between fbtype and time with the between-subject factor or covariate were not significant (all $p > .23$). In Figure 2.4 (p.32), performances in intervention runs (points 2, 4, 6 on the X-axis) are displayed along with the performance in intervening test runs (points 1,3,5,7 on the X-axes) in the actual sequence of their presentation.

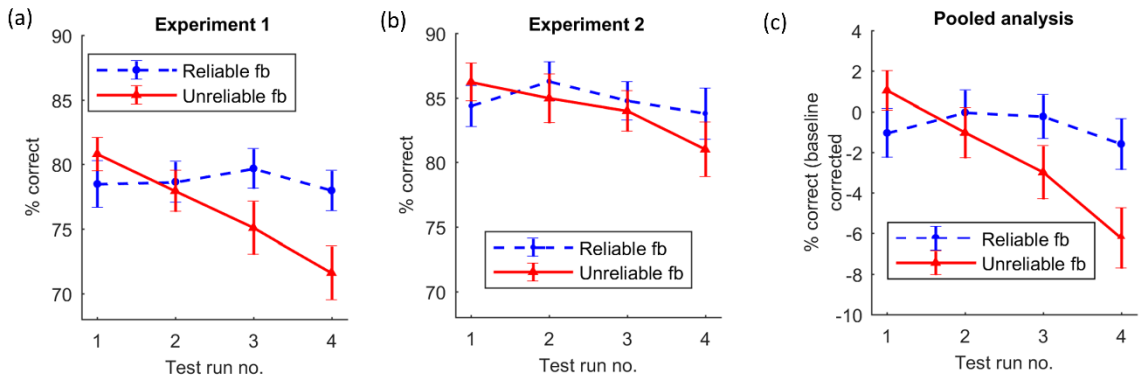

Figure 2.3: Performance in the test runs across time and feedback type in (a) experiment 1, (b) experiment 2 and (c) across the pooled data (for illustrative purposes, data have been corrected for baseline differences in performance between the two experiments). Errorbars show standard errors of the means.

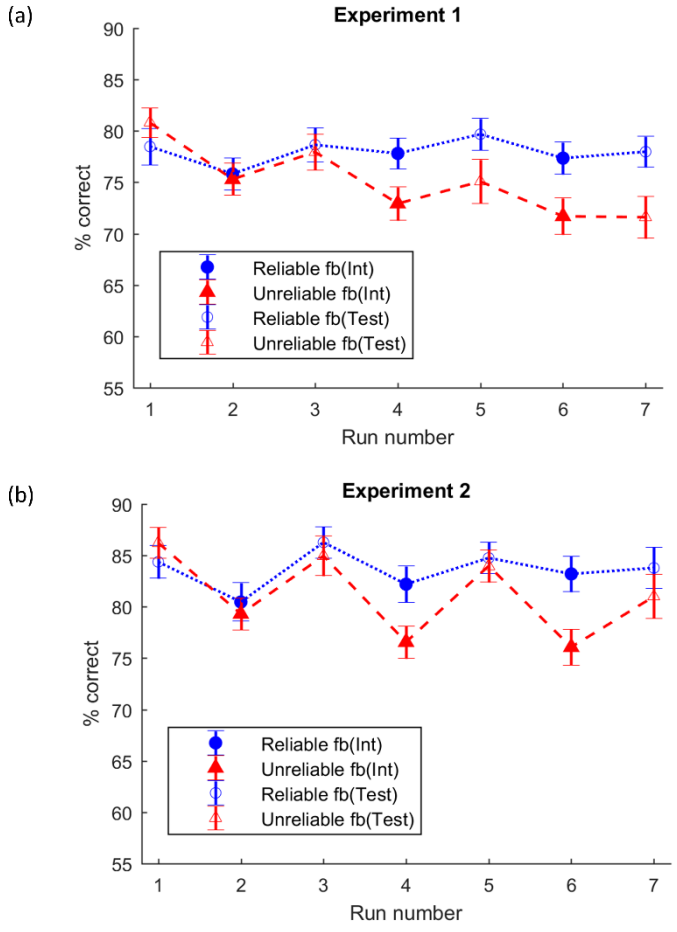

Figure 2.4: Performance in the intervention runs (filled triangles and circles) across time and feedback type in (a) experiment 1 and (b) experiment 2. Test runs are included for comparison (unfilled triangles and circles). Errorbars show standard errors of the means.

Thus, taken together, these results show that unreliable feedback systematically impairs the accuracy of perceptual decision-making, and that this effect was present even after the delivery of such feedback stopped.

2.3.3.2. Unreliable feedback on perceptual decisions shifts responses towards prior beliefs

In order to test our hypothesis that the behavioural responses shift towards prior beliefs, cue-congruence indices (CCI, equation 2.3, p.27) were compared across time (test runs 1 to 4) and fbtype (unreliable/reliable) using the same analysis (two-way RM-ANOVA) that was used to study performance changes. Hence the critical analysis here too was the linear interaction between fbtype and time. This linear interaction effect did not reach the significance threshold in experiment 1 ($F(1, 29) = 2.81$, $p = .1$; Figure 2.5a, p.35), but it did so in experiment 2 ($F(1, 27) = 4.51$, $p = .04$; Figure 2.5b, p.35). However, the session-wise slopes showed that in both the experiments, there was an increase in the CCI across time in the unreliable feedback session as seen by the significant positive slopes (experiment 1: slope $M = 2.76$, $SE = 1.14$, $t(33) = 2.42$, $p = .02$; experiment 2: slope $M = 4.42$, $SE = 1.43$, $t(31) = 3.09$, $p = .004$), but not in the reliable feedback sessions, as seen by the corresponding non-significant slopes (experiment 1: $M = –0.41$, $SE = 1.04$, $t(31) = –0.4$, $p = .69$; experiment 2: $M = 0.63$, $SE = 1.39$, $t(31) = 0.45$, $p = .65$). None of the between-subject factors and covariates interacted significantly with fbtype and time (all $p > 0.08$).

Similar to the pooled analysis that was performed for the performance data in the previous sub-section, we performed a post-hoc RM-ANOVA of the CCI data pooled across experiments 1 and 2 with the additional between-subject factor *experiment number* to study the common mechanisms underlying both experiments. This analysis revealed a significant linear interaction between fbtype and time ($F(1,57) = 6.76$, $p = .01$, Figure 2.5c, p.35), which was based on a positive slope for unreliable feedback ($M = 3.57$, $SE = 0.91$, $t(65) = 3.93$, $p < .001$) and a non-significant slope for reliable feedback ($M = 0.09$, $SE = 0.86$, $t(65) = 0.11$, $p = .91$). Thus, unreliable feedback on perceptual choices increases the reliance on priors when they become available.

Complementary analysis

Since CCI is a difference, it is impossible to conclude whether the observed changes resulted from (1) *enhancements* in performance across time (test runs) in the cue-congruent (CC) trials, (2) *deteriorations* in performance across time in the cue-incongruent (CI) trials, or (3) a combination of (1) and (2). To clarify this, we performed RM-ANOVAs with the same within-subject factors, between-subject factors and between-subject covariates as used in the analyses of overall performance and CCI. The analysis of congruent (CC) trials showed that the interaction between fbtype and time was not significant. Although there was a trend-wise interaction for the CC trials in experiment 1 ($F(1, 29) = 4.01$, $p = .055$, see Figure 2.6a, p.35), this was likely due to the general decrease in the overall performance in the unreliable feedback session in experiment 1 (compare with Figure 2.3a, p.32). Further, there was no fbtype-by-time interaction for the CC trials in experiment 2 ($F(1, 27) = 0.11$, $p = .74$; Figure 2.6b, p.35). The interaction between fbtype and time was significant in the CI trials in both experiments (experiment 1: $F(1, 29) = 7.87$, $p = .009$; experiment 2: $F(1, 27) = 5.51$, $p = .03$; Figure 2.6c-d, p.35), which corresponded to significant (one-sampled t-tests) negative slopes for the unreliable feedback sessions (experiment 1: $M = –5.13$, $SE = 1.13$, $t = –4.51$, $p < .001$; experiment 2: $M = –4.98$, $SE = 1.38$, $t = –3.6$, $p = .001$) , and the absence of any significant slopes in the reliable feedback sessions (experiment 1: $M = –0.27$, $SE = 0.94$, $t = 0.28$, $p = .78$; experiment 2: $M = –0.8$, $SE = 1.22$, $t = –0.66$, $p = .52$). Thus, the increase in CCI observed as a result of unreliable feedback across runs is best explained by the second option, i.e., a greater deterioration in performance in the wrongly predicted (CI) trials.

2.3.3.3. Unreliable feedback decreases confidence in responses during feedback delivery

With the confidence ratings collected in experiment 2, we explored whether unreliable feedback would give rise to lower confidence in perceptual decisions. Confidence was encoded as a decimal value in the interval [0.03, 1], where 0.03 and 1 were the lowest and highest possible ratings of confidence, respectively (see Section 2.3.2.1 on p.26 and Figure 2.2b on p.22). Similar to the earlier analyses, we first tested for a linear interaction between fbtype and time on the confidence ratings in the test runs. Contrary to the results from the objective measures (performance and CCI), no interaction was observed for the

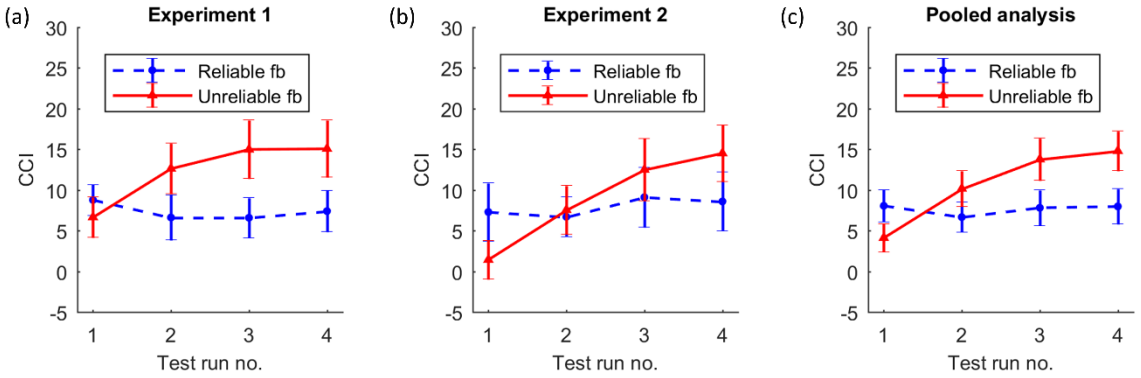

Figure 2.5: Cue-congruence index (CCI) in the test runs across time and feedback type in (a) experiment 1, (b) experiment 2 and (c) across the pooled data. Errorbars show standard errors of the means.

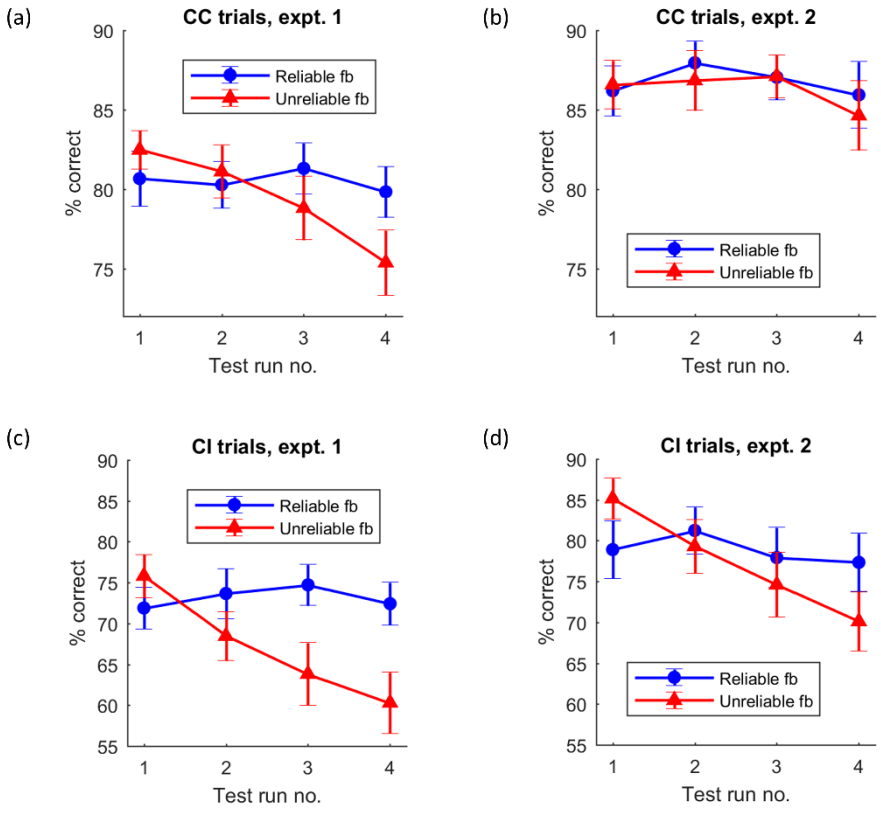

Figure 2.6 : Changes in accuracy in (a-b) congruent and (c-d) incongruent trials across time (test runs 1-4, X-axis) and fbtype (unreliable/reliable) in experiments 1 (a, c) and 2(b, d).

subjective measure, i.e., confidence ($F(1, 27) = 0.84$, $p = .37$, Figure 2.7). Nevertheless, there was a small but highly significant increase in confidence across time in the reliable feedback sessions (slope $M = 0.03$, $SE = 0.01$, $t(31) = 3.12$, $p = .004$), but not in the unreliable feedback sessions (slope $M = -0.002$, $SE = 0.01$, $t(31) = -0.2$, $p = .84$).

Looking at the unreliable feedback sessions (Figure 2.7), there were sharp drops of confidence in the intervention runs of the unreliable feedback session (average confidence: $M = 0.35$, $SE = 0.03$), which relaxed to baseline in subsequent test runs ($M = 0.51$, $SE = 0.04$). A comparison of the mean confidence in the test runs and the mean confidence in the intervention runs (both pooled across time) for the unreliable feedback session showed that the drop in confidence was highly significant ($M = -0.16$, $SE = 0.02$, $t(31) = -6.67$, $p < .001$). There was a slight decrease in mean confidence across the intervention runs compared to test runs even in the control sessions with reliable feedback as well – however, the decrease was much smaller here ($M = -0.06$, $SE = 0.01$, $t(31) = -4.31$, $p < .001$).

Thus, while unreliable feedback has long-lasting effects on objective measures of perceptual inference (i.e., task performance and CCI) that transfer to test runs, it exerts a short-term effect on the subjective measure (confidence), affecting it only *during* the actual intervention.

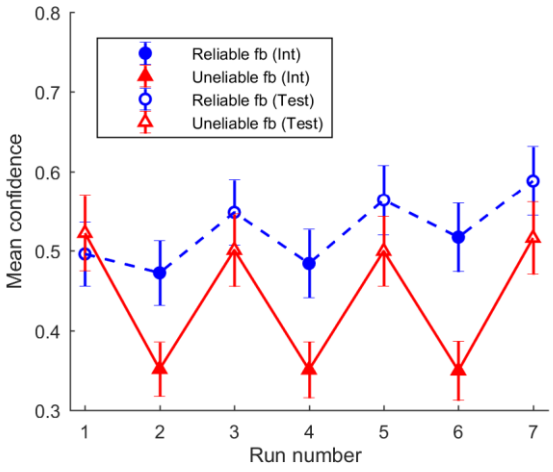

Figure 2.7: Mean confidence in the test runs (unfilled circles and triangles) and intervention runs (filled triangles and circles) across time and feedback type in experiment 2. Errorbars show standard errors of the means.

2.3.3.4. Other post-hoc tests

Some additional tests were conducted to rule out the influence of non-perceptual factors such as awareness of feedback manipulation, motivation and initial position of the mouse cursor. These tests are described below:

(i) Awareness of feedback manipulation

Participants indicated their awareness of feedback manipulation during debriefing at the end of the second session of both the experiments (see Sections 2.3.1.9., p.25 and 2.3.2.4., p.28-29). Only a small fraction of participants (experiment 1: 23.53%, experiment 2: 9.38%) were completely unaware of the feedback manipulation (score 0). Larger percentages of participants were partly (score 0.5, experiment 1: 32.35%, experiment 2: 43.75%) or fully (score 1, experiment 1: 44.12%, experiment 2: 46.88%) aware of the manipulation.

To investigate the relationship between the awareness scores and the main results (i.e., the decrease in performance and the increase in cue-congruence), we performed correlations (Figure 2.8, p.38) between the awareness scores 0, 0.5 and 1, and the slope differences for performance ($\delta_{\text{Perf. slope}}$) and cue-congruence ($\delta_{\text{CCI slope}}$) using Spearman's rank correlation for both performance (experiment 1: $r = .37$, $p = .03$; experiment 2: $r = .03$, $p = .89$; Figure 2.8a-b) and CCI (experiment 1: $r = -.35$, $p = .04$; experiment 2: $r = -.18$, $p = .31$; Figure 2.8c-d).

Thus the results showed that the magnitudes of the performance and cue-congruence effects decreased (i.e., got closer to zero) as the awareness of feedback manipulation increased in experiment 1, but there was no relationship between the effects and awareness in experiment 2. Critically, in experiment 1, the effects did not amplify with increasing awareness of feedback manipulation – if anything, awareness diminished our effects. Thus, the observed changes in performance and cue-congruence effects were not the result of deliberate manipulation by participants.

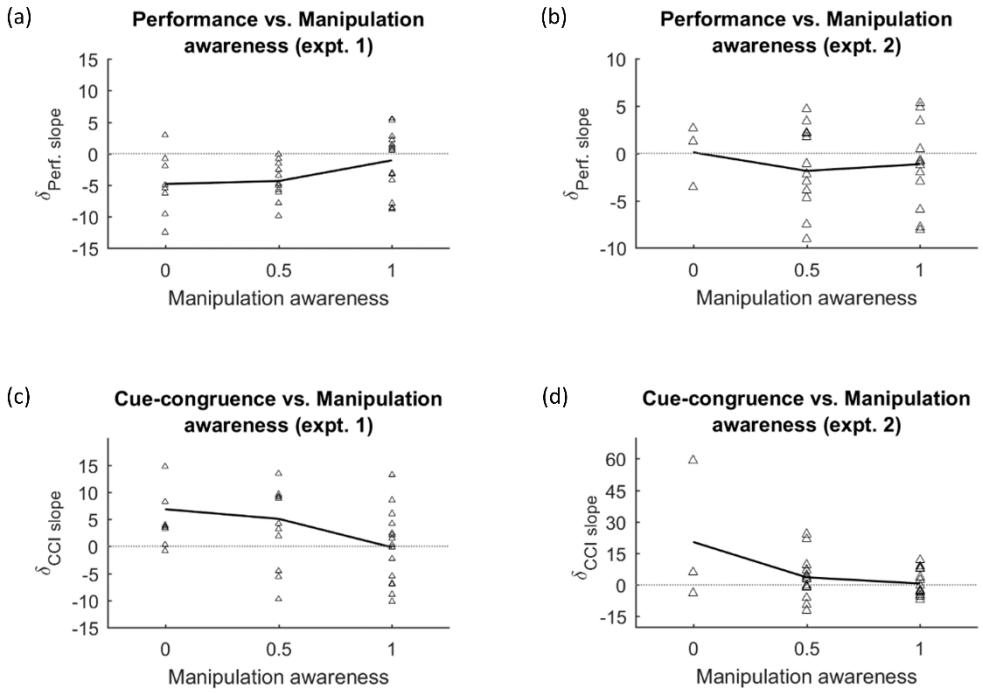

Figure 2.8: Relationships between the awareness of feedback manipulation (X-axes) and the slope differences for performance ($\delta_{\text{Perf. slope}}$, a-b) and cue-congruence ($\delta_{\text{CCI slope}}$, c-d) for experiments 1(a, c) and 2 (b, d). In all plots, the triangles represent individual participants, and the black lines connect the mean slope differences at each level of awareness.

(ii) Changes in motivation (experiment 2 only)

A second post-hoc analysis investigated the influence of motivation on performance and cue-congruence. This analysis was performed only for experiment 2, where motivation ratings were collected for each run (Figure 2.9, p.39). Comparison of session-wise slopes between sessions revealed that there was no difference between motivation ratings on the unreliable and reliable feedback sessions (M = –0.5, SE = 1.28, $t(31)$ = 0.39, p = .7, paired t-test). Next, we tested for correlations between slope differences of motivation ratings ($\delta_{\text{Motiv. slope}}$) and analogous slope differences of the main dependent variables ($\delta_{\text{Perf. slope}}$ and $\delta_{\text{CCI slope}}$). This correlation was not significant, both for performance (r = .05, p = .79, Figure 2.9a) and for CCI (r = –.12, p = .52, Figure 2.9b). Thus it appears that there is no direct influence of motivation on performance and CCI.

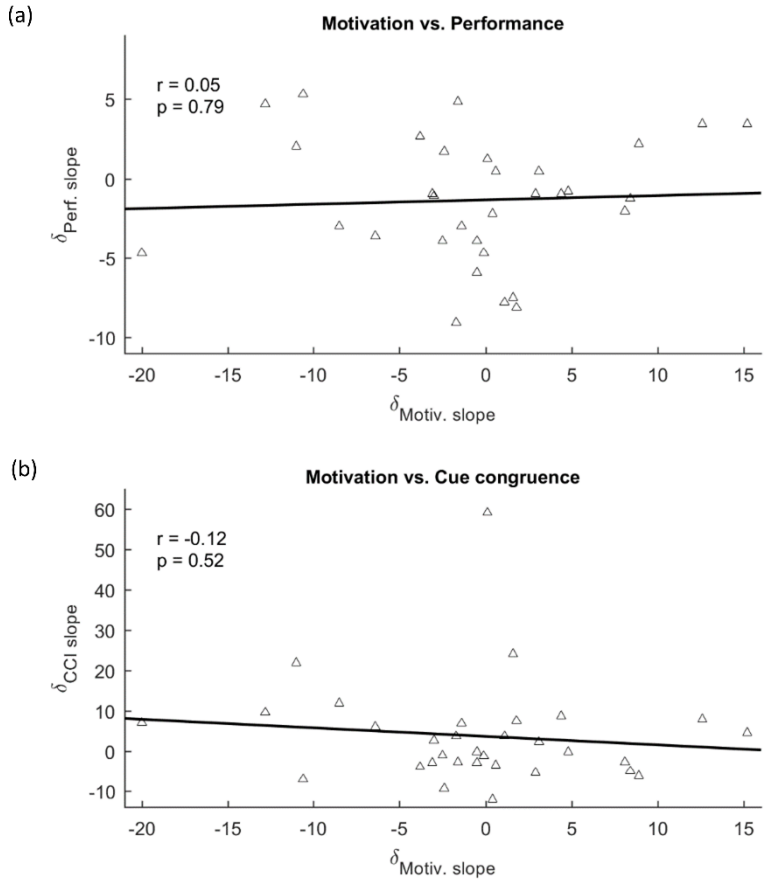

Figure 2.9: The slope difference for motivation ($\delta_{\text{Motiv. slope}}$, X-axis) plotted against the slope differences for (a) task performance ($\delta_{\text{Perf. slope}}$) and (b) cue-congruence ($\delta_{\text{CCI slope}}$) for experiment 2. In both plots, the triangles correspond to individual participants, and the black lines represent the linear fit of the data points.

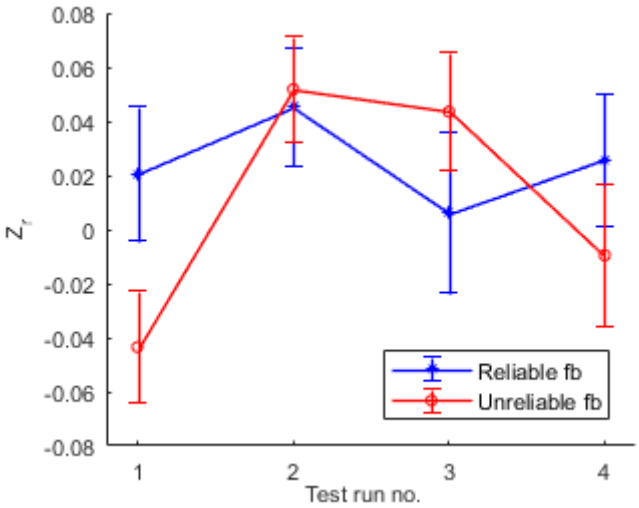

Figure 2.10: The influence of initial cursor position on the responses in experiment 2. Z-transformed correlation coefficients (Z_r, Y-axis) plotted across time (X-axis) and feedback type. Errorbars denote standard error.

(iii) Initial position of the cursor (experiment 2 only)

In the third post-hoc test, we investigated whether the initial random position of the cursor had any influence on participants' responses in experiment 2. There was no correlation overall between the initial cursor position and responses, as seen from the one-sample t-test of the Z-transformed subject-wise correlation coefficients (M = 0.02, SE = 0.01, $t(31)$ = 1.47, p = .15). Further, such correlations also did not emerge during the experiment, as seen from the absence of linear interaction between time and fbtype ($F(1, 27)$ = 0.38, p = .54, Figure 2.10, p.39) and main effect of time ($F(1, 27)$ = 0.95, p = .34). Thus, the randomisation of the initial cursor position did not bias participants' responses.

2.4. Simulation

Simulations were performed to forward-model the effects of unreliable and reliable feedback and to test our hypotheses *in silico* using artificial subjects. This served two purposes: (i) it allowed us to test our hypotheses on a much larger sample than is practically possible in the lab, and (ii) it allowed us to understand possible underlying mechanisms of unreliable feedback on learning and decision-making. The details of the simulation and implementation with artificial subjects are given below.

2.4.1. Procedure

The sensory data were modeled in the form of normal distributions with variance 4 and a mean value of either 0 (N0, "target absent") or 0.5 (N1, "target present"). Variance and mean values were chosen such that they matched the target performance of the staircase procedure in the behavioural experiment (approximately 80% correct).

Training of the observer's sensory classifier was based on two normal-gamma distributions (D0 and D1) that were used to learn mean value and variance of the two classes represented by the distributions N0 and N1. A normal-gamma distribution is a four-parameter distribution, which represents a probabilistic estimate of the moments of a normal distribution and is updated with new samples from this normal distribution (i.e., it is a conjugate prior for normal distributions with unknown mean and variance in Bayesian learning). This means that in the hypothetical case of n $\rightarrow \infty$ correct

observations (infinite samples from the true underlying normal distribution), the normal-gamma distribution would represent a certain estimate of the underlying normal distribution, and the posterior predictive distribution would then converge to the true distribution. The training procedure is shown schematically in Figure 2.11a.

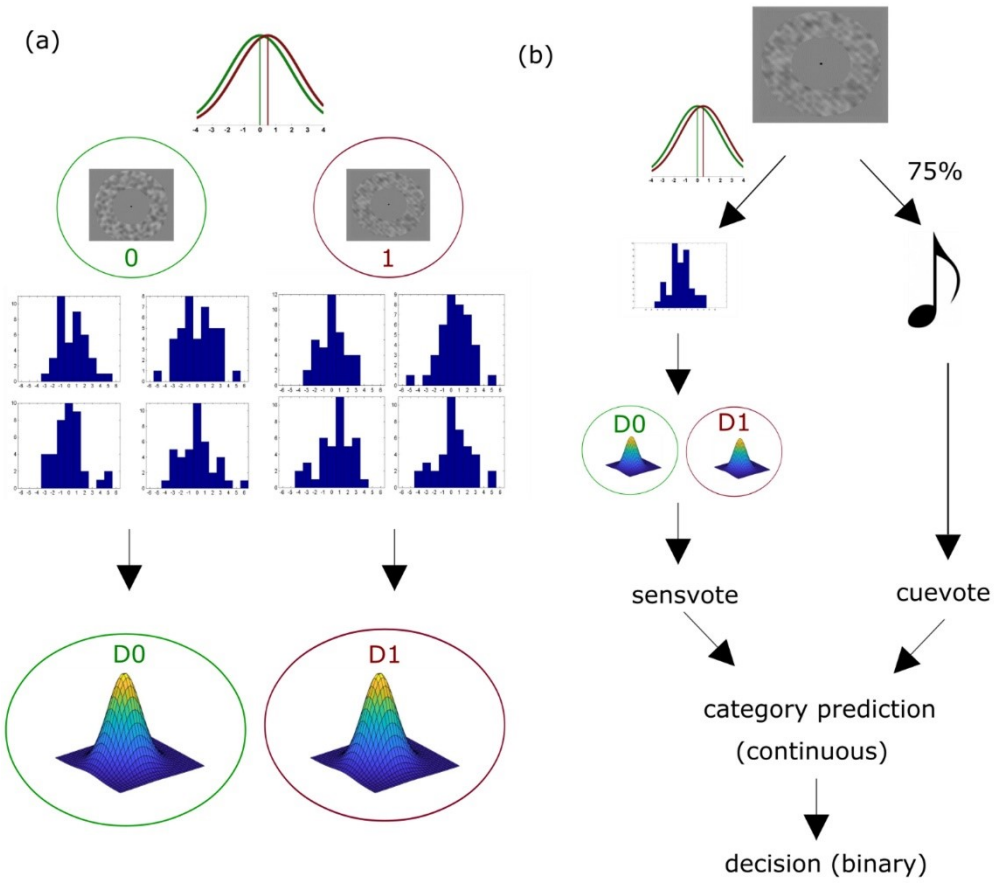

Figure 2.11: Details of the simulation. (a) Training of the sensory classifier in the simulation (training and intervention runs). Two normal-gamma distributions (D0 and D1) were fed with samples (histograms) from two normal distributions '0' ("target absent") and '1' ("target present"). (b) Decision-making in the simulation (test runs). A vote of the sensory classifier about the presence of absence of target (sensvote) was obtained by comparing the likelihoods of the sensory data between D0 and D1. This vote was combined with cue information (cuevote) in a logistic function to make decisions (equations 2.4-2.6).

The likelihood that the new samples belong to one of the stimulus classes represented in the respective normal-gamma (D0 or D1) can be evaluated with the respective posterior predictive distributions (obtained by integrating over the normal-gamma distribution), each of which took the form of a student t-distribution. Hence, the likelihood of the data, given the predictive distribution of the normal-gamma, is a measure of the probability that the data is drawn from the normal distribution (N0 or N1)

represented by the respective normal-gamma distribution. Here, the observer collected 40 samples from the normal distribution in each trial. Next, the vote of the sensory classifier (called *sensvote* here) was obtained by subtracting the log likelihood of the data given hypothesis 0 (samples are drawn from the normal distribution N0 represented in D0) from the log likelihood of the data given hypothesis 1 (samples are drawn from the normal distribution N1 represented in D1). Mathematically, this can be represented as follows:

$$sensvote = ln\,(p(X|D1)) - ln\,(p(X|D0)) \qquad (2.4)$$

The sensory vote thus obtained is a measure of how much more likely the sensory data originates from the "target present" distribution N1 as compared to the "target absent" distribution N0. Thus, *sensvote* is a single value obtained for each trial, without an associated distribution or precision. Similarly, a trialwise *cuevote* was obtained as binary values 0 or 1, encoding "target absent" and "target present", respectively. To identify their individual contributions to the actual stimulus (i.e., target absent/present), the regression coefficients corresponding to sensvote and cuevote (β_s and β_c, respectively) were obtained by fitting these terms to the *actual* stimulus category using logistic regression. This is because at this stage, the simulated observer learns from reliable feedback providing information about true stimulus categories, analogous to learning from reliable feedback in the test runs of the behavioural experiments. The estimated regression coefficients are in theory comparable to precisions (or inverse variances) of distributions corresponding to different sources of information. Next, to estimate the behavioural outcome, the decision classifier took the form of a logistic regression with *sensvote* and *cuevote* as predictors, and then converted the predictions to binary decisions as shown below in equations 2.5-2.6:

$$prediction = 1 + \frac{1}{1+\exp(c+\beta s*sensvote+\beta c*cuevote)} \qquad (2.5)$$

$$decision = \begin{matrix} 0 & prediction < 0.5 \\ 1 & prediction \geq 0.5 \end{matrix} \qquad (2.6)$$

Predictions below and above 0.5 were assigned to the categories "target absent" and "target present", respectively. The constant *c* in equation 2.5 was estimated using maximum likelihood optimisation (implemented in the fitglm routine of the Matlab Statistics and Machine Learning Toolbox) and was included to improve the flexibility of the model in case of unequal apriori probabilities of stimuli. The decision-making procedure is shown schematically in Figure 2.11b (p.41). The simulation used functions from the Statistics and Machine Learning Toolbox of MATLAB.

2.4.2. Implementation

To mimic the above-chance performance of human observers at baseline, the sensory classifiers (distributions D0 and D1) were pre-trained with 20 samples of "stimuli" (N0 and N1). This was followed by simulations of test and intervention runs alternating in a manner similar to the behavioural experiments (Figure 2.1c, p.20). In the test runs, stimuli were classified by the decision classifier based on sensory data and the cue information as described above. As the goal of the simulation was specifically to investigate the effect of learning from unreliable feedback, the sensory classifiers were not updated in the test runs, i.e., there was no learning in the simulated test runs. In the intervention runs, the sensory classifiers (distributions D0 and D1) were updated with samples from distributions N0 and N1. If the feedback was reliable, distribution D0 was always updated with samples from N0 and D1 with samples from N1. However, if the feedback was unreliable, each distribution (D0 and D1) was trained with N0 in one half of the trials and N1 in the other half, i.e., half of the virtual stimuli were mislabelled (analogous to invalid feedback trials in the behavioural experiments). The number of virtual trials, or updates, were identical to those of the behavioural experiments – 64 trials in the simulated test runs and 128 trials in the simulated intervention runs.

The main hypotheses about the effects of unreliable feedback (relative to reliable feedback) on task performance and cue-congruence were tested using the simulated data. 1000 iterations were performed each of unreliable and reliable feedback interventions. These were taken to be 1000 artificial "subjects" in data analysis. Based on the predicted responses, performance and CCI were computed for each test run (see Section 2.3.2.2., p.27), and the effect of unreliable feedback was computed using a two-way RM-ANOVA with time (test run number) and fbtype (unreliable/reliable feedback) as within-subject

factors. Significant interactions were further explored using session-wise slopes, computed for each artificial subject by linearly fitting run-wise performance and CCI data across time (test runs) separately for unreliable and reliable feedback.

2.4.3. Results

The simulated data was tested in the same manner as the data from the behavioural experiments, i.e., using two-way RM-ANOVAs with time (test run number) and fbtype (unreliable/reliable) as factors. We note that the simulation-based values for performance and CCI cannot be interpreted in absolute terms as these depend on arbitrary simulation parameters representing the initial moments of the two stimulus distributions. Thus, only *changes* in performance and CCI can be inferred from the simulations.

2.4.3.1. Unreliable feedback impairs task performance in the simulated participants

The linear interaction between fbtype (reliable/ unreliable) and time (test runs 1 to 4) was significant ($F(1, 999) = 92.42$, $p < .001$) for the simulated data, in line with our hypothesis (Figure 2.12a, p.45). Post-hoc analyses showed that the observed interaction was based on a decrease in performance, indicated by the significantly negative slope across time in the unreliable feedback session ($M = -0.8$, $SE = 0.07$, $t(999) = -12.32$, $p < .001$) and a non-significant slope across time in the reliable feedback session ($M = 0.01$, $SE = 0.06$, $t(999) = 0.14$, $p = .89$). Thus, the simulation further verified our hypothesis that unreliable feedback impairs performance in perceptual decision making.

2.4.3.2. Unreliable feedback enhances cue-congruent responses in the simulated participants

In line with our hypothesis and the behavioural data, the simulation showed a significant linear interaction effect for CCI ($F(1,999) = 166.7$, $p < .001$; Figure 2.12b, p.45). Further in line with our second hypothesis, CCI increased over time in the unreliable feedback

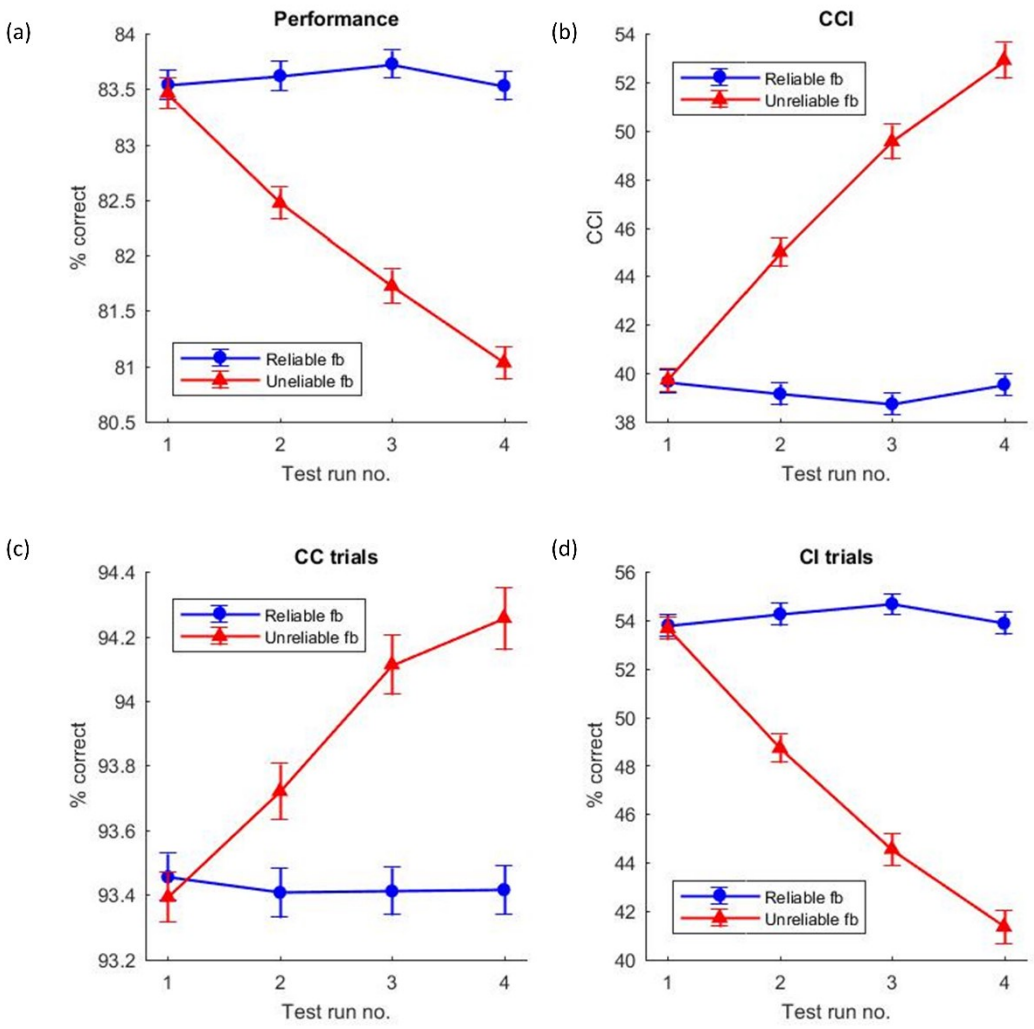

Figure 2.12: Results from the simulated data. (a) Task performance, (b) Cue-congruence index (CCI), (c) performance in congruent trials (CC) and (d) performance in incongruent trials (CI) across fbtype and time for the simulated data. Note that the Y-axis values are vastly different across (a), (c) and (d). Errorbars show standard errors of the means.

session, resulting in a positive slope (M = 4.41, SE = 0.29, $t(999)$ = 15.06, p < .001), and did not change in the reliable feedback session, showing no change in slope (M = –0.08, SE = 0.2, $t(999)$ = –0.39, p = .69).

Next, to understand the changes in performance in the CC and CI trials, we computed run-wise performances separately for CC and CI trials, and analysed changes across time and fbtype using two RM-ANOVAs. The analysis revealed that there was a significant interaction between fbtype and time in the CC trials ($F(1, 999)$ = 36.83, p < .001, Figure 2.12c), which was revealed to be due to an increase in performance across

time as indicated by the positive slope in the unreliable (M = 0.3, SE = 0.04, t(999) = 7.59, p < .001) and not the reliable (M = –0.01, SE = 0.03, t(999) = –0.34, p = .73) feedback session. Similarly, a significant interaction was also observed for the CI trials (F(1, 999) = 156.75, p < .001, Figure 2.12d, p.45), which was shown to result from a decrease in performance across time as indicated by the negative slope in the unreliable feedback group (M = –4.11, SE = 0.28, t(999) = –14.84, p < .001) but not in the reliable feedback group (M = 0.07, SE = 0.2, t(999) = 0.34, p = .74). Thus, the simulations confirm that unreliable feedback would result in an increase in CCI, contributed both by improved task performance in congruent trials and decreased task performance in the incongruent trials.

2.5. Discussion

This study investigated the influence of unreliable feedback on perceptual inference using two behavioural experiments and a simulation. We observed that when unreliable feedback was provided to perceptual decisions, (1) task performance deteriorated, and (2) perceptual inference was shifted towards learned priors in the following period.

Previous studies have already shown trends in the direction of our observations – that unreliable feedback prevents learning (Herzog & Fahle, 1997, 1999), changes the sensitivity to stimuli (Aberg & Herzog, 2012) and that it induces false percepts even after delivery of such feedback stops (Vannucci et al., 2011; Whitson & Galinsky, 2008). Further, studies have also shown that a decrease in the precision of likelihood distributions can shift the inference towards priors (Körding & Wolpert, 2004). Since the prior (defined by a fixed cue-stimulus association) in our experiments did not interact with unreliable feedback across the course of the experiment, the observed increase in cue-congruent responses was likely due to the decline in accuracy in the processing of sensory information. Please note that if unreliable feedback would have been presented in the presence of predictive cues or priors, the precisions of both the priors and the sensory evidence would likely have been affected, which may have reduced or even nullified the observed prior-congruent behaviour. Here, in order to delineate the influence of unreliable feedback on the processing of the bottom-up sensory evidence and to prevent a direct learning between cues and unreliable feedback, we kept the prior separate from the feedback manipulation. This is also more compatible with real-world

perception, where priors are learnt over the course of life and are hence unlikely to change due to temporary uncertainty.

The decrease in performance was stronger in experiment 1 than in experiment 2, the latter experiment resulting in a negative performance slope for the unreliable feedback session, but no significant linear interaction between fbtype and time. This difference between the two experiments could be due to several reasons. One possibility is that the baseline performance (test run 1) was higher in experiment 2 than experiment 1, thereby reducing task difficulty and consequently the disruptive influence of unreliable feedback. The task could have also been less perceptually demanding, since more time was given to make responses on the continuous scale in experiment 2. Finally, the effect of unreliable feedback may simply differ between detection and discrimination tasks. Different neural mechanisms have been proposed to underlie detection and discrimination, both in terms of the neurons encoding visual stimuli at lower levels (Hol & Treue, 2001) and in the cognitive resources required to perform the tasks (Sagi & Julesz, 1984). Discrimination has been suggested to be more demanding and to involve two subsets of neurons instead of one. However, since we see the opposite effect in our results, this is an unlikely explanation for the differences seen between the two behavioural experiments.

Conceivably, non-perceptual mechanisms may have led to the observed increase in prior congruence. For instance, following the cue could have been a conscious strategy adopted by participants. However in such cases, the accuracy in congruent (CC) trials would have drastically increased, and the performance in CI trials would have reduced drastically, approaching 100% and 0%, respectively. However, from Figure 2.6 (p.35), we can see that in the unreliable feedback sessions, (1) performance in CC trials does not increase and (2) that the decrease in performance in CI trials is of a smaller magnitude. This pattern of results suggests a slow and rather automatic shift in responses towards the prior.

We also performed post-hoc tests to identify potential confounds in our results due to participants' awareness of the feedback manipulation, subjective motivation or the initial position of the cursor (when a continuous response scale was used). Results revealed that the effects ($\delta_{\text{Perf. slope}}$ and $\delta_{\text{CCI slope}}$) did not increase with a higher awareness of feedback manipulation. In fact, experiment 1 showed that higher awareness of feedback

manipulation reduced the differences between sessions and thereby the efficacy of feedback manipulation. Changes in motivation did not correlate with changes in performance and cue-congruence either. Lastly, in experiment 2, the initial position of the cursor on the response bar did not influence the responses. These results attest that the performance and cue-congruence effects did not arise from deliberate strategies or differences in subjective motivation. However, we acknowledge that participants' awareness of unreliable feedback could have reduced the sizes of our effects.

Since confidence is an indicator of performance in addition to metacognition (Guggenmos et al., 2016; Hebart et al., 2016; Yeung & Summerfield, 2012) , it is possible to suggest that in experiment 2, confidence simply mirrors performance. However, it must be noted that although the relative differences between sessions (unreliable vs. reliable) were similar between performance and confidence, the actual events are slightly different: unreliable feedback *prevents an increase* of confidence whereas it *decreases* performance accuracy. The influence of unreliable feedback on performance and confidence should be tested in a future experiment where reliable feedback is not delivered in the test runs, which might help to counteract the immediate restoration of performance and confidence.

Taken together, the simulations and the experiments detailed in this chapter suggest that unreliable feedback, when given to perceptual choices, has a debilitating effect on performance and skews perception towards prior beliefs. While this study could not determine whether these effects stemmed from changes in sensory processing in the visual cortex or due to changes in higher-level decision-making processes (Herzog & Fahle, 1999; Rahnev, Nee, Riddle, Larson, & D'Esposito, 2016), Study II, described in the next chapter, used neuroimaging to investigate the neural processes underlying the effects of unreliable feedback on perceptual inference.

3. Study II: Unreliable Feedback Deteriorates Information Processing in the Primary Visual Cortex

The work presented here has been submitted as:

- Varrier, R. S., Rothkirch, M., Stuke, H., Guggenmos, M., & Sterzer, P. Unreliable feedback deteriorates information processing in primary visual cortex.

3.1. Introduction

As detailed in the preceding chapters, perception is an inferential process, whereby an internal model of the world is used to infer the most probable causes of the sensory data (Friston, 2005; O'Reilly et al., 2012). For such perceptual inference to be adaptive, the reliability of the sources of sensory data must be taken into account: Highly reliable sensory information should be given more weight in perceptual inference than unreliable information (Adams et al., 2013; Knill & Pouget, 2004). Typically, reliability of the sensory information is manipulated in experiments by adding varying levels of noise to stimuli, and it has been shown that this results in neural stimulus representations in sensory areas that are less informative (Darcy et al., 2019; Hebart et al., 2012; Ludwig et al., 2016). We hypothesised that neural representations may not only be affected by the currently available sensory information, but also by learned beliefs regarding its reliability. To test this, we designed a functional magnetic resonance imaging (fMRI) experiment in which sensory stimulation in a challenging visual discrimination task was kept constant, while a belief about the uncertainty of sensory information was induced by giving unreliable feedback on task performance. We predicted that this manipulation would lead to a deterioration of neural stimulus representations.

Previous studies have shown that the delivery of unreliable feedback in visual and auditory tasks impaired performance and prevented perceptual learning (Herzog & Fahle, 1997, 1999; Vuvan et al., 2018). Further, participants had a higher tendency to see patterns in noise (Whitson & Galinsky, 2008) and at lower signal-to-noise ratios (Vannucci et al., 2011). Noisy feedback was also shown to shown to shift responses away from sensory information and towards prior knowledge in a visuo-motor task (Körding & Wolpert, 2004). These previous reports strongly suggest that sensory information gets down-weighed under conditions of environmental uncertainty; however the neural changes underlying such behaviour has not been studied yet.

Based on previous fMRI studies showing successful decoding of visual grating orientations from activation patterns in primary visual cortex (V1) (Kamitani & Tong, 2005; Haynes & Rees, 2005; Kok et al., 2012), we employed an orientation discrimination task and examined changes in V1 using pattern distinctness (Allefeld & Haynes, 2014), an index that estimates the dissimilarity between multivariate activity patterns of competing stimuli. The participants had to perform a challenging visual orientation discrimination

task whose difficulty was determined by the deviation of each stimulus from a reference orientation in a clockwise (CW) or counter-clockwise (CCW) direction. There were two such reference orientations (45° and 135°) and correspondingly two pairs of stimuli (Figure 3.1, p.53). In order to rule out the role of motor preparation during stimulus presentation, the stimulus-response mapping was also randomised. Thus, participants were informed of which response buttons to use to indicate their percepts only after the stimuli disappeared (Hebart et al., 2012; Kahnt, Grueschow, Speck, & Haynes, 2011).

Critically, the stimulus orientations were determined for each participant prior to the main experiment based on their individual performance thresholds, and thus the bottom-up sensory information remained the same all through the main experiment. Similar to Study I, beliefs about uncertainty were induced by providing unreliable feedback on performance in a dedicated intervention phase, and the effects of such interventions on sensory representations of stimuli were measured by neural activity patterns in the test phases that precedes and followed the intervention phase. To minimise learning during the test phases, feedback was withheld in these runs. To control for some of the general effects of the task such as stimulus exposure, motivation, attention, fatigue etc., a separate group of age- and gender-matched participants performed the same experiment, but with reliable feedback in the intervention phase instead of unreliable feedback. A between-subject design was adopted to maximise uncertainty and to reduce the possibility of participants detecting external feedback manipulation.

3.2. Hypothesis

In this study, we hypothesised that relative to reliable feedback, unreliable feedback would lead to (1) a decrease in task performance accuracy in a visual orientation discrimination task and (2) a decrease in the distinctness of multivariate patterns corresponding to grating stimuli in V1.

3.3. Materials and methods

3.3.1. General design

Each participant took part in one experimental session, where they were first trained in the task outside the fMRI scanner, following which they were taken to the scanner, where a threshold estimation was performed first (when brain images were not acquired), followed by the main experiment (where functional images were acquired during the task). In the main experiment, stimuli were presented at the orientations determined in the preceding threshold estimation step. The main experiment consisted of a pre-intervention test phase without feedback, an intervention phase with either unreliable or reliable feedback and a post-intervention test phase without feedback (Figure 3.1a, p.53). Thus, the study had a between-subject design, where participants were assigned to one of two experimental groups that differed only with respect to the feedback delivery in the intervention phase – one group received trial-wise feedback on task performance that was valid at chance-level (in 50% of the trials), whereas the other group received trial-wise performance feedback that was always valid (in 100% of the trials).

3.3.2. Participants

The study was approved by the ethics committee at Charité - Universitätsmedizin Berlin, and informed consents were collected from all participants. Participants were students from Humboldt University and Charité – Universitätsmedizin Berlin. Thirty-two healthy participants took part in the experiment (ages 18-35, mean age = 24.9, 13 female). Of these, two participants were excluded from the experiment before the intervention phase due to chance-level performance in several runs in the pre-intervention phase. This left us with 30 participants – with 15 participants each in the unreliable and reliable feedback groups. FMRI data from a participant in the reliable feedback group had to be discarded due to excessive head motion during the post-intervention phase. Further, the fMRI data from two (out of eight) runs in the post-intervention phase of eight participants (N=4 from each group) were lost due to an error in the scanner sequence. In these participants, the corresponding two runs from the pre-intervention runs were likewise excluded for fMRI data analysis to make the pre- and post-intervention data comparable in terms of statistical power for the multivariate analyses.

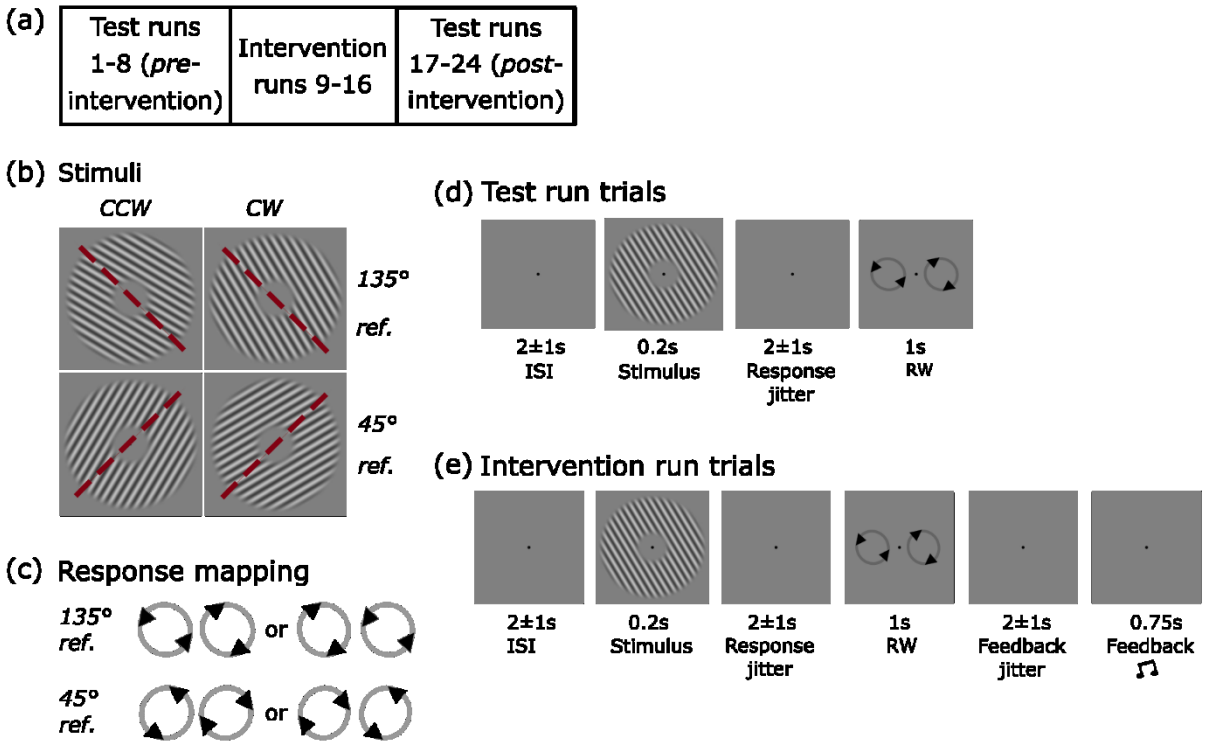

Figure 3.1. (a) Time course of the main experiment (i.e., after threshold estimation). (b) Sinusoidal grating stimuli used for the orientation discrimination task. There were two pairs of stimuli corresponding to the implicit diagonals at 135° and 45°. Reference lines are shown in red here for illustrative purposes, they were not presented during the actual experiment. The depicted orientation deviations from diagonals are also exaggerated for illustration. (c) On each trial, one of two types of response mapping was assigned randomly: CCW (CW) responses were to be indicated by pressing either the left (right) button or the right (left) button of a response box. Time courses of trials in (d) test runs and (e) intervention runs. The response cues are enlarged in (d) and (e) for better visibility. ISI = inter-stimulus interval, RW = response window.

3.3.3. Stimuli

Stimulus presentation was implemented using PsychToolbox 3.0.11 (psychtoolbox.org) for Matlab (MathWorks Inc.). Visual stimuli were presented on a monitor (resolution: 1024 x 768 pixels) and projected using an oblique mirror into the eyes of participants lying in a supine position (total distance 154 cm). The stimuli were high-contrast annular sinusoidal gratings (inner radius = 1.32°, outer radius = 6.69°, spatial frequency = 1.29cpd) with luminance ranging from 25% to 75% of the maximum luminance of the screen. To reduce visual responses to the sharp inner and outer edges of the annular

stimuli against the grey background, these edges were blurred using circular Gaussian functions centred at the inner (1.38°) and outer (6.64°) edges such that the contrast gradually fades until it matches the grayscale background (standard deviation σ = 0.44°). This procedure rendered high contrast sinusoidal gratings with soft edges (see Figure 3.1b, p.53). Image properties such as contrast, spatial frequency and size were not changed between various orientations and across the whole experiment. To reduce neural adaptation and to avoid "point-of-reference" strategies by the participants, the Gabor patches were presented with variable phase shifts in different trials, randomly drawn from 16 equally spaced shifts between 0 and 2π.

On each trial, one of four stimuli could be presented (Figure 3.1b, p.53) – i.e., stimuli rotated clockwise (CW) or counter-clockwise (CCW) relative to the two reference orientations, i.e., 45° or 135°. In the training phase (described in Section 3.3.7.1, p.56), participants learnt to use the relevant diagonal for each type of stimulus. The diagonal to be compared against was the one which was in the same quadrant as the presented stimulus (for example, if the presented stimulus had an orientation of 60°, the correct percept would be that it was clockwise, since the 45° diagonal was the correct reference to use). The degree of deviation from the diagonal references was determined in the threshold estimation phase prior to the main experiment, so that the same four stimulus orientations were presented all through the main experiment.

3.3.4. Responses

In order to orthogonalise stimuli and responses and to prevent motor planning during stimulus presentation, participants were informed of the stimulus-response mapping (i.e., which button to press to indicate the perception of the CCW and CW orientations) only after the stimulus disappeared (Hebart et al., 2012; Kahnt et al., 2011). All possible sequences of the response cues that were presented are illustrated in Figure 3.1c (p.53). Response cues were presented for a time window of 1s, during which participants were asked to make responses (Figure 3.1d-e, p.53). The cues were small circles with arrows indicating the response mapping (CCW/ CW) and were presented on the left and right sides of the fixation dot with their inner arcs at 0.48° visual angle. There were two pairs of response cues corresponding to stimuli with 45° and 135° references, and these could be presented in two sequences (left-right or right-left). Participants familiarised

themselves with the response cues and how to use them during the initial training outside the scanner.

3.3.5. Feedback

Trial-by-trial feedback was delivered by means of auditory tones after each response during training, threshold estimation and in the intervention phase of the main experiment (see Section 3.3.7., p.56-57). The tones were positive, negative or neutral, depending on whether the response was correct, incorrect or missed, respectively. Participants in the unreliable feedback group received pseudo-randomised feedback in the intervention runs, such that in half of the trials, the feedback delivered was faulty (i.e., positive tones after incorrect button presses and negative tones after correct button presses). Participants in the reliable feedback group always received valid feedback. Similar to the stimuli and response cues, participants also familiarised themselves with the feedback tones and their associated meanings during the initial training period.

3.3.6. Trials

The time course of test and intervention trials are shown in Figure 3.1d-e (p.53). On each trial, participants' task was to report their perceived orientation of the grating as either CCW or CW with respect to an implicit diagonal reference (see Figure 3.1b on p.53 and Section 3.3.3. on p.53-54). Each trial started with the presentation of a fixation dot (radius 0.1° visual angle) for 2±1s, followed by the presentation of the visual stimulus for 0.2s, which was followed by fixation for 2±1s. Next, the response window was presented for 1s, during which the response mapping was indicated using the response cues, and responses were made by pressing one of two buttons on an fMRI-compatible button box. In the intervention phase alone, following the response, there was another fixation window (2±1s), following which the auditory feedback was delivered (0.75s). On each trial, the durations of the fixation windows (ISI, response and feedback) were sampled randomly from a distribution consisting of 8 evenly spaced values between 1 and 3 (1.0000, 1.2857, 1.5714, 1.8571, 2.1429, 2.4286, 2.7143, 3.0000).

3.3.7. Experimental schedule

At the beginning of the experiment and outside of the scanner, participants were trained in the orientation discrimination task using supra-threshold versions of the gratings (20 minutes). Once inside the scanner, individual orientation discrimination thresholds for the main experiment were determined using a staircase procedure (10 minutes). 100% valid feedback was delivered both in the training and the threshold estimation steps to facilitate learning of the task and the response mapping. The main experiment consisted of 24 runs (overall 85 minutes), and this was followed by a short functional localiser task (6 minutes). At the end of the experiment, participants were debriefed and compensated.

3.3.7.1. Training

Participants performed training runs in a testing room outside the scanner. The first run consisted of supra-threshold stimuli and the participants manually navigated through each stimulus and response screen at their own pace. The second run consisted likewise of supra-threshold stimuli, but trial timings corresponded to those of the main experiment (time course shown in Figure 3.1e, p.53). If necessary, the second run was repeated until the participant could make responses in the given time and got at least 80% correct responses.

3.3.7.2. Staircase procedure

Inside the scanner, participants performed a staircase task to set the threshold deviation of stimuli from the diagonal references at which they could discriminate between CCW and CW orientations with moderate difficulty. To this end, we used a two-down, one-up staircases with equal step-sizes up and down which arrived at an 80% performance threshold. The first phase of the staircase procedure determined the approximate signal threshold and had larger step-sizes (angular deviations from the two diagonals were multiplied by the factors $10^{-0.03}$ and $10^{0.03}$ to decrease or increase thresholds, respectively). The second phase started at the threshold estimated by the first staircase and used a fixed step size of 0.3°, both for an increase and for a decrease in deviations from the diagonal. Each staircase stopped when a certain number of reversals (six

reversals for phase one, ten reversals for phase two) were reached or 80 trials were completed. Thresholds were estimated by averaging the last four and six reversal points for the staircase phases one and two, respectively. On average, this procedure resulted in comparable threshold deviations across the two participant groups (unreliable: M = 7.85°, SE = 1.53°, reliable: M = 8.08°, SE = 0.89°; two-tailed, two-sample t-test: $t(28) = 0.13$, $p = .9$).

3.3.7.3. Main experiment

The main experiment consisted of 24 runs, split into three parts (Figure 3.1a, p.53): Runs 1-8 were pre-intervention test runs (without feedback), runs 9-16 were intervention runs (with feedback) and runs 17-24 were post-intervention test runs (without feedback). Each run consisted of 32 trials, in which each of the four types of stimuli (threshold CCW/CW deviations from the 45° and 135° diagonal references) were shown 8 times. In the intervention runs, half of the participants received reliable feedback, and the other half received unreliable feedback. The pre- and post-intervention test runs were identical in structure, and their purpose was to measure changes induced during the intervention phases with reliable/unreliable feedback.

3.3.7.4. Functional localiser

A functional localiser run was included after the main experiment, in which the four stimulus conditions of the main experiment were presented, along together with a fifth fixation-only baseline condition. Each condition was shown in a block of 12-second duration, and the conditions were repeated six times in a pseudo-random order. During the 12-second presentations of the four stimuli, the 16 phase shifted visual stimuli (mentioned in sub-section Stimuli, Section 3.3.3., p.53-54) were presented randomly at a rate of 3.33Hz. To ensure that participants fixated during the functional localiser run, a central fixation dot was present in all conditions and changed its colour to red briefly (0.3s) at random, and participants were asked to press the left response button to indicate whenever this event occurred.

3.3.7.5. Debriefing

At the end of the experiment, all participants were given questionnaires to probe their awareness of having received unreliable feedback and motivation to do the task. Participants also rated their motivation to do the task in each phase of the experiment as a percentage value. The relevant questions are given below (verbatim):

(i) How reliable(correct) was the feedback (0 = absolutely not reliable(correct),100 = totally reliable(correct))

0%---------- 100%

(ii) During the experiment, did you think that the feedback was manipulated? Please circle your response

Definitely yes/ highly likely / maybe / highly unlikely/ definitely not

(iii) Did your trust in the feedback change at any point during the experiment? If so, please mention at what point approximately that occurred.

(iv) On a scale of 0(absolutely not motivated) to 100(fully motivated), how motivated where you during

(a) Run 1: Blocks 1 to 8, no feedback

(b) Run 2: Blocks 1 to 8, with feedback

(c) Run 3: Blocks 1 to 8, no feedback

(d) Run 4: 1 block, colour change detection task

Please note that in this context, "run" and "block" allude to what we refer to in the rest of the chapter as a "phase" and "run", respectively.

3.3.8. Eye-tracking

To ensure fixation, an MRI-compatible video-based eye-tracker (iView XTM MRI 50Hz, SensoMotoric Instruments, Teltow, Germany) was used to monitor participants' gaze position throughout the experiment. Eye-tracking data could not be collected from two participants due to difficulties in calibration or in the detection of pupil and corneal reflex by the camera. In all other participants, partial or full data were collected and pre-processed using the following steps: (1) removal of invalid data points, (2) cubic-spline

interpolation of missing data points when there was fewer than eight missing data points (160ms), (3) removal of linear trends and (4) computation of running averages across five consecutive points (100ms). After pre-processing, data corresponding to the stimulus presentation windows (200ms) were extracted. Next, participants with a high proportion of missing data (defined as (i) more than 70% invalid data points overall or within the stimulus presentation windows of the pre- or the post-intervention phases, or (ii) no data at all from more than six out of the eight runs present within each pre- or the post-intervention phase). The area inside the inner edge of the stimuli was selected as the fixation window (radius 1.32°), and fixation performance was quantified for each test phase as the percentage of data points that were within this fixation window during stimulus presentation. Overall, valid data was retrieved successfully from 22 participants (12 from the unreliable feedback group and 10 from the reliable feedback group).

3.3.9. FMRI data acquisition and processing

3.3.9.1. Data acquisition

Functional brain images were acquired at a 3T Siemens Trio (Erlangen, Germany) scanner using a gradient echo-planar imaging sequence and a 12-channel head coil. Each run in the pre- and post-intervention phases consisted of 90 T2*-weighted whole-brain volumes each, and the functional localiser consisted of 180 whole-brain volumes. Other parameters remained the same during the main experiment and the localiser (TR = 2s, TE = 30ms, flip angle = 78°, 33 slices, descending acquisition, 3mm isotropic resolution, 0.7mm gap between slices). In addition, high-resolution structural T1-weighted images were acquired using the MPRAGE sequence (TR = 1.9s, TE = 2.52ms, flip angle = 9°, 192 slices, 1mm isotropic resolution).

3.3.9.2. FMRI data processing

The functional images were corrected for slice acquisition delays and translational/rotational motion using the MATLAB-based Statistical Parametric Mapping Toolbox (SPM12, www.fil.ion.ucl.ac.uk/spm). Next, the functional images in their native space were co-registered with the structural image of the single-subject brain template

Colin27 (which had been inverse-normalised to the native subject-space) to correct for a misalignment between them – this was important for later anatomical voxel selection (since the anatomical mask was in alignment with the Colin27 mask). After this step, the functional images were smoothed with a 3mm (FWHM) Gaussian kernel, since recent studies support spatial smoothing in multivariate analysis (Gardumi et al., 2016; Hendriks, Daniels, Pegado, & Op de Beeck, 2017; Misaki, Luh, & Bandettini, 2013; Op de Beeck, 2010). Next, three general linear models (GLMs) were defined for each participant corresponding to the pre-intervention runs, the post-intervention runs and the functional localiser runs. In each GLM, the four stimuli (Figure 3.1b, p.53) were included as separate regressors, and a fifth regressor encoded the response windows with button presses. These regressors were then convolved with the canonical haemodynamic response function as implemented in SPM12. In addition, the six translation and rotation parameters obtained from the motion correction step were included in each model as regressors of no interest. The GLMs from the pre- and post-intervention runs were used to estimate pattern distinctness (see below), and the GLM from the functional localiser was used to create a T-contrast map of voxels that responded to the visual field region in which the four visual stimuli were presented (Stimulus > Fixation), which was used for the selection of voxels for the estimation of pattern distinctness. Thus, voxels within brain area V1 were selected if they (1) had a probability of greater than 50% of belonging to V1 (area hOc1 in the SPM-based Anatomy Toolbox (Eickhoff et al., 2005)) and (2) were significant at an uncorrected T-contrast threshold of 0.05. Additionally, T-contrast maps (all stimuli > 0) were also computed for each test phase (pre-/ post-intervention) to later estimate mean activity within the V1 mask for a complementary analysis.

Estimation of pattern distinctness: To estimate the distinctness of stimulus-evoked BOLD activation patterns in V1, we used the cross-validated (CV) MANOVA algorithm (Allefeld & Haynes, 2014). CV-MANOVA performs a leave-one-run-out cross-validation to compute an unbiased estimate of the distinctness of activation patterns. When used to compare *two* multivariate patterns, this measure of pattern distinctness is analogous to the Mahalanobis distance. In the current experiment, the pattern distinctness thus corresponded to the cross-validated Mahalanobis distance between stimulus pairs with the same diagonal reference (i.e., pair 1 consists of gratings deviating CW and CCW from the 45° reference, and pair 2 consists of gratings deviating CW/CCW from the 135°

reference, see Figure 3.1b, p.53). Higher values of pattern distinctness indicate more dissimilarity between competing stimuli, and correspondingly, better stimulus representations. By averaging across the pattern distinctness for the two stimulus pairs, we obtained a single estimate of pattern distinctness for each participant and test phase (pre-/post- intervention).

3.4. Statistical analyses

The key behavioural dependent variable was the orientation discrimination performance during the pre- and post-intervention test phases, which was quantified as the percentage of correct responses within each phase. The key dependent variable for the fMRI data analysis was the mean pattern distinctness obtained from each test phase as described in the previous section.

2x2 mixed-design ANOVAs consisting of the between-subject factor feedback type (fbtype: unreliable or reliable) and the within-subject factor test phase (time: pre-/post-intervention) were performed separately for the two dependent variables, namely, task performance and pattern distinctness. The orientation discrimination threshold was included as a covariate of no interest in both the analyses. Our critical prediction was that there would be a significant interaction between the factors fbtype and time as a result of the relative decrease in performance and pattern distinctness in the unreliable feedback group. In case of significant interactions, post-hoc one-sample t-tests (two-tailed) were performed of the changes in performance ($\Delta_{\text{Percent correct}}$) and pattern distinctness ($\Delta_{\text{Pattern distinctness}}$), computed as differences between post- and pre-intervention values. Thus, $\Delta<0$, $\Delta=0$ and $\Delta>0$ would correspond to deteriorations, absence of changes and enhancements, respectively, in performance/ pattern distinctness. To determine the effect sizes of the behavioural and neural changes, Cohen's d corrected for the small sample size (Durlak, 2009) was estimated separately for changes in performance and pattern distinctness between the two groups (unreliable/ reliable feedback).

Since the pattern distinctness at each test phase was averaged between stimulus pairs that had a common diagonal (i.e., pair 1: CW/CCW deviations from the 45° diagonal and pair 2: CW/CCW deviations from the 135° diagonal, see Figure 3.1b, p.53), we additionally tested for differences in $\Delta_{\text{Pattern distinctness}}$ between the two stimulus pairs using

separately for the reliable and unreliable feedback groups using paired t-tests to identify potential biases in our results.

Next, to test if the changes in pattern distinctness induced by the delivery of unreliable feedback paralleled similar changes in performance, Pearson correlation was computed between the changes in performance ($\Delta_{\text{Percent correct}}$) and pattern distinctness ($\Delta_{\text{Pattern distinctness}}$). In order to test for the robustness of this correlation, this correlation was additionally computed after the removal of outliers. Outliers were defined as the $\Delta_{\text{Percent correct}}$ or $\Delta_{\text{Pattern distinctness}}$ data points that deviated from the first or third quartile of the dataset by more than 1.5 times of the inter-quartile range (IQR). This correlation analysis was performed only for the participant group that received unreliable feedback, since our hypothesis was about the effect of *unreliable* feedback on behavioural responses and neural representations, and not about the effects of feedback in general.

To examine potential changes in attention (Kastner et al., 1999, 1998) as a result of unreliable feedback, a complementary 2x2 mixed-design ANOVA of the overall V1 activity (within the voxel mask defined in Section 3.3.9.2. on p.59-60) was performed using the same factors (fbtype and time) and covariate as were used in the analyses of task performance and pattern distinctness. The dependent variable here was the BOLD activity during stimulus presentation (obtained from the T-contrast "all stimuli > 0"; see Section 3.3.9.2. on p.59-60 for more) averaged across the task-relevant V1 voxels.

Lastly, a few control analyses were performed. First of all, to rule out the possibility that the differences in performance and pattern representations between the two groups could be driven by differences in fixation, a 2x2 mixed-design ANOVA was performed on the eye-tracking data with the same factors and covariate as used in the main analyses of performance and pattern distinctness. The dependent variable for this analysis was computed for each test phase as the percentage of data points that were within the fixation window during stimulus presentation. For more details, please refer to Section 3.3.8 (p.58-59). Next, the responses to the debriefing questions (Section 3.3.7.5., p. 58) were analysed. Answers to questions about the awareness of feedback manipulation were compared between the two groups (unreliable/ reliable feedback) and correlated with the changes in task performance ($\Delta_{\text{Percent correct}}$) and pattern distinctness ($\Delta_{\text{Pattern distinctness}}$). Further, to examine whether motivation varies in a manner similar to that of performance and pattern distinctness, a 2x2 mixed-design ANOVA was performed using motivation

ratings as the dependent variable and using the same factors (fbtype and time) and covariate (orientation discrimination threshold) as used to study performance and pattern distinctness. Last of all, performance of the colour-detection task during the functional localiser run was compared between the two groups, to see if there was a general, task-independent decrease in performance in the unreliable feedback group compared to the reliable feedback group, and to verify that participants fixated well during the functional localiser (this was critical for proper voxel selection).

3.5. Results

3.5.1. Unreliable feedback impairs task performance

The delivery of unreliable feedback in the intervention phase led to a significant decline in discrimination performance, as shown in a two-way mixed-design ANOVA, where the between-subject factor feedback type (unreliable vs. reliable) and the within-subject factor time (pre- vs. post-intervention) showed a significant interaction effect ($F(1, 27) = 7.26$, p = .01; Figure 3.2, p.64). Post-hoc two-tailed one-sample t-tests showed a significant decrease in performance after unreliable feedback ($M = -6.88$, $SE = 2.60$, $t(14) = -2.64$, $p = .02$) but not reliable feedback ($M = 3.49$, $SE = 2.73$, $t(14) = 1.28$, $p = .22$). The effect size (Cohen's d) of the change in performance between the two groups was 0.94.

3.5.2. Unreliable feedback deteriorates stimulus representations in V1

Neural effects of unreliable feedback were assessed by estimating the distinctness of activation patterns in stimulus-responsive voxels of V1 evoked by CW- vs. CCW- rotated gratings using cv-MANOVA (Allefeld & Haynes, 2014). FMRI data were available from 29 participants (unreliable feedback: n = 15, reliable feedback n = 14; see Section 3.3.2., p.52 for details). The voxel selection process (described in Section 3.3.9.2., p.59-60) yielded comparable numbers of V1 voxels in both groups (unreliable feedback: 225.4 ± 13.51, reliable feedback: 219.25 ± 17.35 voxels; two-tailed, two-sample t-test: $t(27) = 0.28$, $p = .78$).

In line with our hypothesis, and in striking analogy to the behavioural results, we found a significant interaction of feedback type and time ($F(1, 26) = 5.98$, $p = .02$; Figure

3.3). Again, post-hoc one-sample t-tests (two-tailed) showed that there was a significant decrease in pattern distinctness after unreliable feedback (M = –0.04, SE = 0.02, $t(14)$ = –2.61, p = .02), but not after reliable feedback (M = 0.02, SE = 0.02, $t(13)$ = 0.98, p = .35).

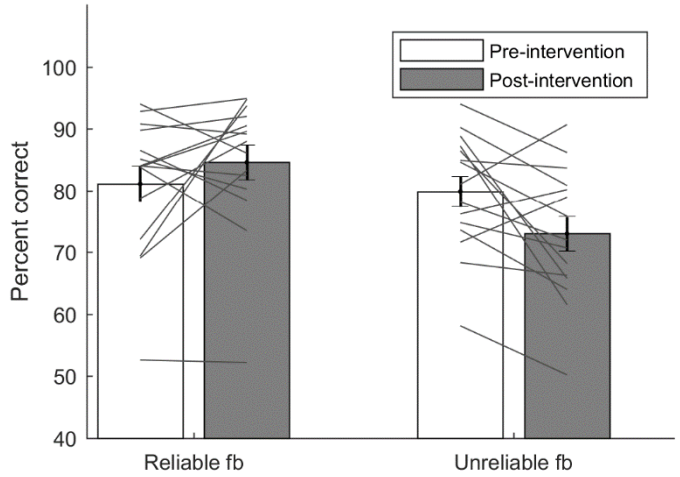

Figure 3.2: Behavioural performance across groups (reliable/unreliable feedback) and test phases (pre-/post-intervention). The bars show mean performances and errorbars show standard errors of the means. The lines show performance accuracies for each participant.

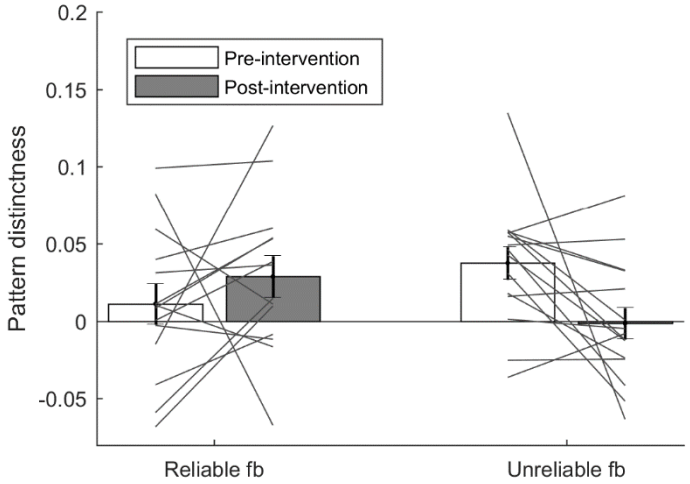

Figure 3.3: Neural pattern distinctness in V1 plotted across groups (reliable/unreliable feedback) and test phases (pre-/post-intervention). The bars show the mean pattern distinctness and errorbars show standard errors of the means. The lines show pattern distinctness estimates for each participant.

The effect size of the change in pattern distinctness between in the two groups was 0.84. We note that pattern distinctness appears to differ between groups at baseline (pre-intervention phase), but this difference was not statistically significant ($t(27) = 1.58$, $p = .13$, two-tailed two-sample t-test). Please note that like the Mahalanobis distance, the *true* pattern distinctness value can never be below zero; however, as stated by the creators of cv-MANOVA, the *estimated* pattern distinctness derived from finite amounts of data can often generate negative values of pattern distinctness when the true value is zero or close to it.

The above results were computed based on mean pattern distinctness obtained from each participant and test phase by averaging the estimates across stimulus pairs with a common diagonal. To verify that these results were not driven by one stimulus pair alone (which would suggest a bias in the results), we used paired t-tests to compare the changes in pattern distinctness ($\Delta_{\text{Pattern distinctness}}$) computed separately for stimulus pairs 1 and 2. The results revealed that $\Delta_{\text{Pattern distinctness}}$ was comparable across the stimulus pairs – for both the group that received unreliable feedback ($t(14) = 0.89$, $p = 0.39$; *45° reference:* $M = -0.05$, $SE = 0.01$; *135° reference:* $M = -0.03$, $SE = 0.02$) and the group that received reliable feedback (t (13) = 1.31, $p = 0.21$; 45° reference: $M = -0.003$, $SE = 0.03$; *135° reference*: $M = 0.04$, $SE = 0.02$).

Since we predicted that unreliable feedback was what led to both the deterioration in performance and multivariate representations, we tested for a correlation between the changes in the two variables ($\Delta_{\text{Percent correct}}$ and $\Delta_{\text{Pattern distinctness}}$) in the unreliable feedback group using a Pearson correlation analysis (Figure 3.4, p.66). In line with our hypothesis, we found a significant positive correlation ($r = .66$, $p = .008$, $n = 15$) between changes in performance ($\Delta_{\text{Percent correct}}$) and pattern distinctness ($\Delta_{\text{Pattern distinctness}}$) after unreliable feedback. This correlation remained significant after removing two outliers ($r = .60$, $p = .03$, $n = 13$), which were identified as the data points deviating from the lower and the upper quartiles by more than 1.5*IQR for either $\Delta_{\text{Percent correct}}$ or $\Delta_{\text{Pattern distinctness}}$.

Complementary analysis

A 2x2 mixed-effects ANOVA performed to evaluate overall activity changes in V1 across fbtype (unreliable/ reliable) and time (pre-/post-intervention) showed that the interaction between fbtype and time was not significant ($F(1, 26) = 0.26$, $p = .62$; Figure

3.5, p.66). Further, the main effects of feedback type ($F(1, 26) = 0.85$, $p = .36$) and time ($F(1, 26) = 0.39$, $p = .54$) were not significant either. Thus, the observed effect on pattern distinctness was not associated with a change in overall neural responsiveness in V1, thus making it unlikely that the deterioration in multivariate representation in V1 resulted from changes in overall attention.

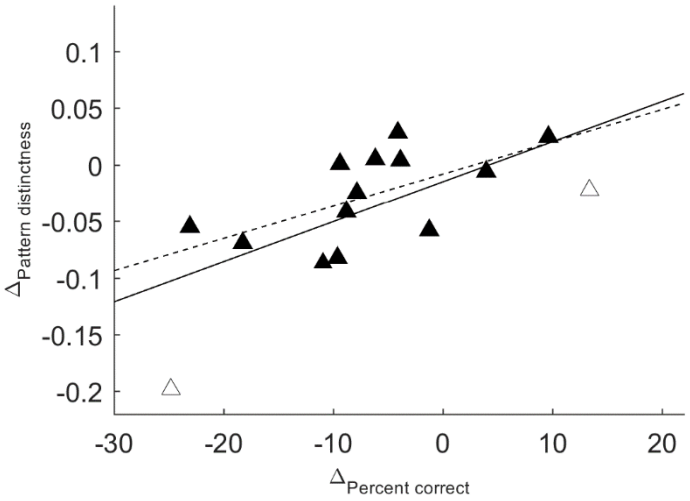

Figure 3.4. Correlation between the changes in performance ($\Delta_{\text{Percent correct}}$) and pattern distinctness ($\Delta_{\text{Pattern distinctness}}$) for the unreliable feedback group. The triangles represent individual participants, and the solid and dashed lines represent regression lines before and after removing the outliers (unfilled triangles), respectively. A positive or a negative value on each axis corresponds to an increase or a decrease as a result of the feedback intervention, respectively.

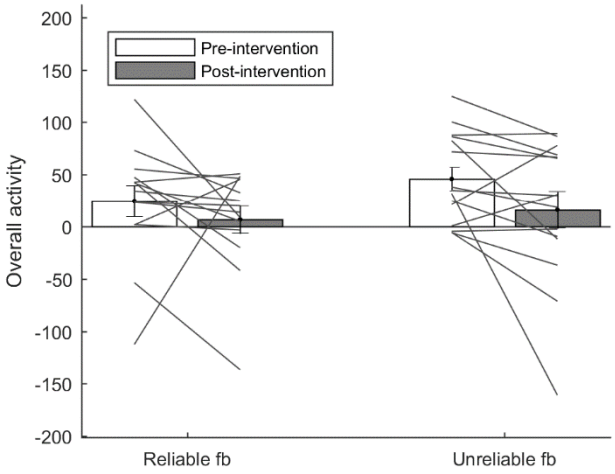

Figure 3.5: Overall brain activity plotted across groups (reliable/unreliable feedback) and test phase (pre-/post-intervention). The bars show the overall brain activity average across participants, and errorbars show standard errors of the mean. The lines show subject-wise estimates.

3.5.3. Control analyses

3.5.3.1. Fixation accuracy did not differ between the groups due to unreliable feedback

We analysed the eye-tracking data to check if the observed effects on performance and pattern distinctness could have risen from differences in fixation across time (pre-/post-intervention) between the two groups (unreliable/ reliable). Data were available from n = 22 participants, and fixation accuracies were determined as the percentage of eye positions within a circular region of interest (radius = 1.3° visual angle) around the fixation dot. Fixation performances were high in general and (in percent, see Figure 3.6) were comparable between the unreliable (pre: M = 88.68, SE = 3.43, post: M = 85.09, SE = 5.51) and the reliable (pre: M = 89.41, SE = 3.00, post: M = 86.25, SE = 9.55) feedback groups. A two-way ANOVA with the factors feedback type and time showed neither significant main effects (fbtype: $F(1, 19) = 0.08$, $p = .78$; time: $F(1, 19) = 0.29$, $p = .60$) nor a significant interaction effect ($F(1, 19) = 0.01$, $p = .94$).

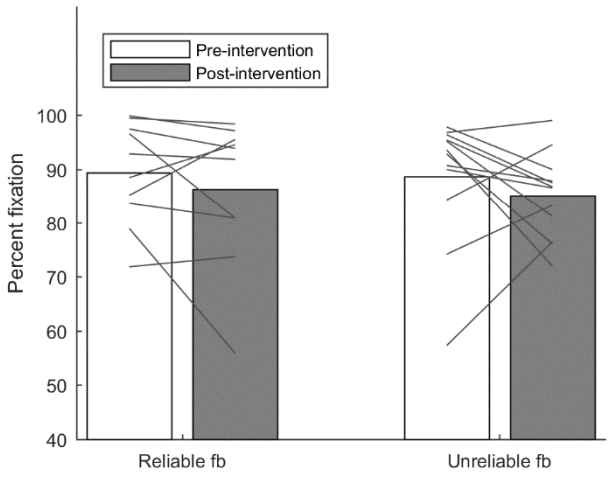

Figure 3.6: Mean fixation percentage plotted across groups (reliable/unreliable feedback) and test phase (pre-/post-intervention). The bars show the mean fixation and the lines show subject-wise estimates.

3.5.3.2. Awareness of feedback manipulation

To study the influence of the awareness of feedback manipulation on the main results, the debriefing data were analysed (Figure 3.7, p.68). Participants who received unreliable

feedback rated a lower percentage reliability (in answer to question (i) in the Section 3.3.7.5., p.58) on feedback (M = 53.67, SE = 1.43, df = 14) than those who received reliable feedback (M = 90.83, SE = 0.58, df = 14), and the difference between the groups was significant (two-tailed, two-sample t-test: $t(28) = 6.2$, $p < .001$, Figure 3.7a). Further, the answer to the question on whether they suspected the feedback to be manipulated was converted to a scale of 1 to 5 (where 1 is the most unaware and 5 is the most aware; qn, (ii) in Section 3.3.7.5., p.58), participants in the unreliable feedback group rated a higher degree of awareness to feedback manipulation (M = 3.8, SE = 0.07, df = 14) than the reliable feedback group (M = 2.33, SE = 0.07, df = 14), and the group-wise difference was significant (two-tailed, two-sample t-test: $t(28) = 3.77$, $p < .001$), Figure 3.7b). Lastly, in the unreliable feedback group, 13 participants reported the intervention run number (between 1 and 8) at which their trust in feedback changed (qn. (iii) in Section 3.3.7.5., p.58), the average of which lay close to the middle of the intervention phase (M = 3.73, SE = 0.51, Figure 3.7c).

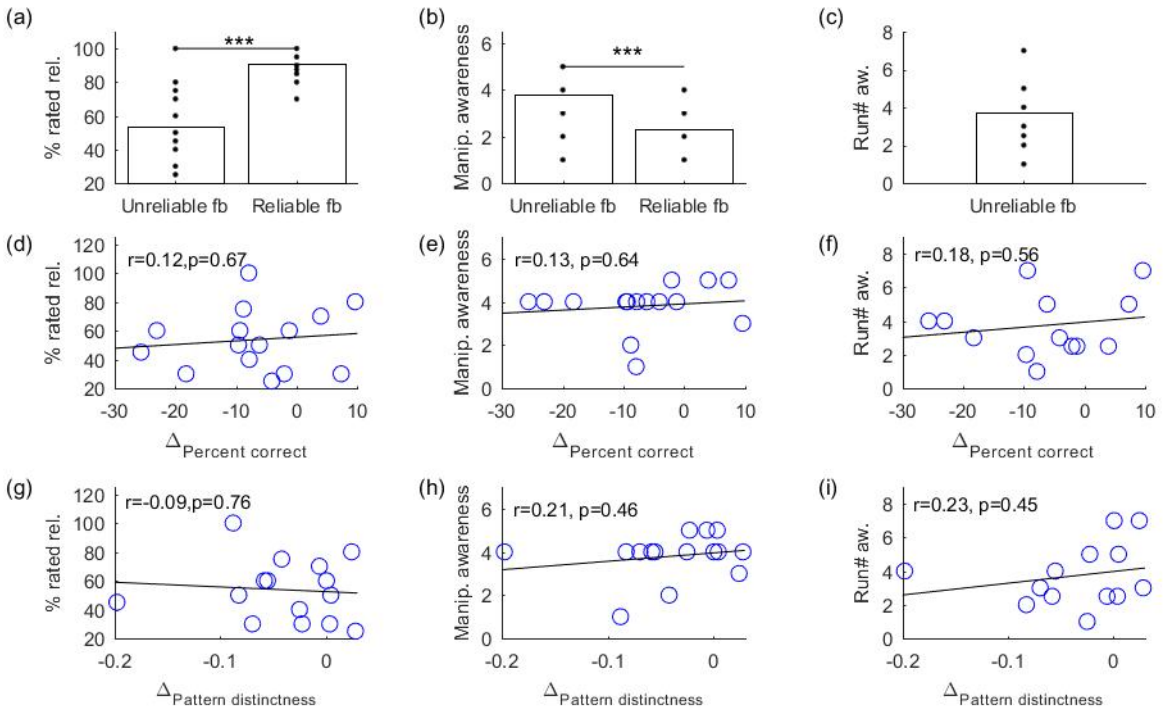

Figure 3.7: (a) Rated reliability (in percent) on feedback, (b) Reported awareness on feedback manipulation converted to integer values 1-5 and (c) intervention run number (1-8) in which feedback manipulation was detected. The relationship between these variables and (d-f) changes in task performance ($\Delta_{\text{Percent correct}}$) and (g-i) changes in pattern distinctness ($\Delta_{\text{Pattern distinctness}}$) for the unreliable feedback group are shown in the middle and lower rows, where each circle represents a participant. Errorbars represent standard errors of the mean, *** indicates p < .001.

To examine if the observed results awareness could have indeed influenced the observed behavioural and neural changes, each of these measures were correlated with $\Delta_{\text{Percent correct}}$ within the unreliable feedback group. We found that there were no significant correlations between $\Delta_{\text{Percent correct}}$ and the three measures of participants' knowledge of external manipulation, namely the percentage reliability of feedback ($r(13) = .12$, $p = .67$, Figure 3.7d), the feedback manipulation awareness ($r(13) = .13$, $p = .64$, Figure 3.7e) and how early the manipulation was detected (performance: $r(11) = .18$, $p = .56$, Figure 3.7f). Similarly, there was no correlation between $\Delta_{\text{Pattern distinctness}}$ and the same three measures – the respective correlations were: percentage reliability ($r(13) = -.09$, p = .76, Figure 3.7g), feedback manipulation awareness ($r(13) = .21$, $p = .46$, Figure 3.7h) and how early the manipulation was detected ($r(11) = .23$, $p = .45$, Figure 3.7i). The lack of correlation between the subjective ratings of awareness with the behavioural and neural effects suggests that the observed effects of unreliable feedback is robust to the awareness of feedback manipulation. Alternatively, participants could have overestimated their awareness of feedback manipulation on being specifically asked about it during debriefing.

3.5.3.3. Motivation ratings did not differ between groups

The 2x2 mixed-design ANOVA to test for changes in the percentage rating of motivation (answer to qn. (iv) in Section 3.3.7.5., p.58) across the factors time and feedback type revealed that neither the interaction effect ($F(1, 27) = 0.41$, $p = .53$, Figure 3.8, p.70) not the main effect of feedback type ($F(1, 27) = 1.59$, $p = .22$) was significant. However, there was a main effect of time ($F(1, 27) = 4.81$, $p = .04$) explained by the overall decrease in motivation from the pre-intervention runs ($M = 86$, $SE = 2.1$) to the post-intervention runs ($M = 69.63$, $SE = 3.43$). Thus, differences in subjective motivation is unlikely to have interfered with the task performance during the main experiment.

3.5.3.4. Performance in functional localiser task was comparable between groups

After the main experiment, there was a functional localiser run in order to select the V1 voxels that respond best to the visual stimuli. Fixation was critical for this, since the visual cortex is retinotopic. To investigate if fixation, and in turn voxel selection, could have been

poorer in the unreliable feedback group, thereby causing the observed differences in the pattern distinctness, we analysed performance accuracies in the colour-change detection task during the localiser run (Figure 3.9). The analysis revealed that task performances were comparable between the unreliable (M = 97.86, SE = 0.3, df = 13) and the reliable (M = 95.33, SE = 0.61, df = 14) feedback groups. Note that one participant reported not having performed the colour-change detection task during debriefing and was hence excluded from this post-hoc analysis.

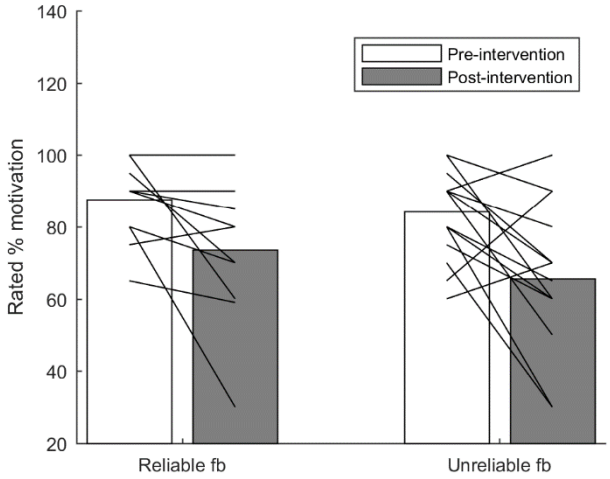

Figure 3.8: Motivation ratings in the pre- and post-intervention phases for the unreliable and reliable feedback groups. The bars show the mean motivations in each test phase and group, and the lines show subject-wise motivations.

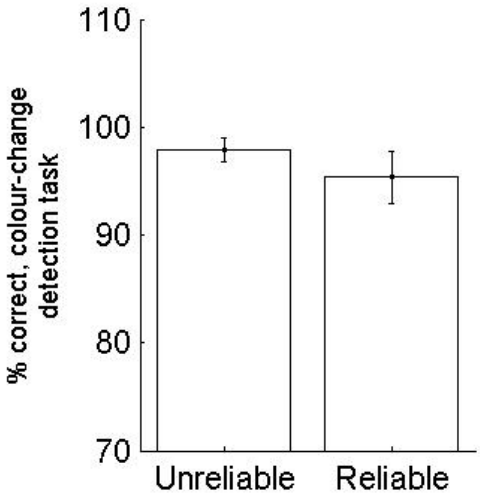

Figure 3.9: Mean performance in the colour-change detection task performed during the functional localiser in the unreliable and reliable feedback groups. Errorbars show standard errors of the mean within each group.

This indicates that the voxel selection procedure was not noisier in the unreliable feedback group and also that the debilitating effects of unreliable feedback on performance could be task-specific.

3.6. Discussion

Taken together, the results from this study that the induction of a belief about uncertainty using unreliable feedback is associated with a deterioration of perceptual task performance and concomitant changes in the neural representation of stimulus information in V1.

While the observed decline in task performance is in line with previous work on unreliable feedback (Herzog & Fahle, 1997, 1999; Vuvan et al., 2018), the neural mechanisms had not been studied before. Here we show that in parallel with the deterioration in performance, there is also a deterioration in the representation of the corresponding grating stimuli in V1. Moreover, individuals with larger performance impairments after unreliable feedback also showed larger drops in V1 pattern distinctness, and vice versa. According to predictive coding theories of hierarchical cortical processing, bottom-up sensory prediction error signals are weighted by the precision of these signals in superficial cortical layers (Adams et al., 2013). In line with this idea, some fMRI studies have also reported activity linked to precision-weighted prediction errors in sensory areas of the brain (Iglesias et al., 2013; Stefanics, Stephan, & Heinzle, 2019). The precision, encoded by the post-synaptic gain of the neurons transmitting the bottom-up signals, is thought to be under the influence of top-down projections (Kanai, Komura, Shipp, & Friston, 2015). We suggest that our finding of reduced distinctness of fMRI signal patterns may be due to a decrease in the precision-weighting of sensory information in V1, most likely mediated by top-down signalling of learned beliefs regarding the reliability of the sensory information.

Two limitations of our study are (1) the small sample size, given the between-group design and (2) the overall high awareness of feedback manipulation during debriefing. Since this study was novel in its design and approach, a formal sample size calculation could not be performed. Further, the between-group design of this study was adopted primarily to prevent participants from noticing the feedback manipulation –

which was reported by a few participants of Study I. In spite of this, participants who received unreliable feedback did report noticeably less reliability in it during debriefing. But since these reliability ratings did not correlate with the observed changes in performance or pattern distinctness at all, it is unlikely that they could have directly influenced our main results. However, to overcome these limitations, future studies should perform such experiments with a larger sample and at a higher task difficulty – although it is to be noted that task difficulty comes at the price of the smaller multivariate effects – since as the task difficulty increases, the physical properties of stimuli (and consequently their neural representations) are likely to get more similar to each other, which in turn would reduce their multivariate decodability and dissimilarity. For now, we recommend that our results be treated as proof-of concept.

Thus, using unreliable feedback as a tool to induce top-down beliefs about uncertainty, we show that the weighing of sensory information in perceptual inference is implemented at the earliest stages of cortical sensory processing and that this can be rapidly changed in accord with current beliefs regarding environmental reliability.

4. General Discussion

4.1. Summary of findings

This thesis set out to investigate the mechanisms by which feedback reliability influences perception and behaviour. We used Bayesian inference theories to accomplish this. We proposed that delivering unreliable feedback (that had a chance-level validity) to perceptual decisions would lead to the down-weighing of sensory information. As a result, we hypothesised that there would be a decline in task performance and an increased reliance on prior beliefs. Further, we predicted that unreliable feedback would decrease the distinctness of stimulus representation in sensory areas of the brain. As a secondary hypothesis, we also predicted that the metacognitive awareness about performance would decrease along with performance. Lastly, we modelled the effects of unreliable feedback on performance and reliance on prior beliefs using a Bayesian learning scheme and logistic regression.

The hypotheses were tested in two empirical studies, described in Chapters 2 and 3 of the thesis. The studies comprised three experiments and one simulation in total. Both the studies showed evidence in favour of our predictions. The findings are summarised below:

(1) **Decrease in task performance** – In both the studies, unreliable feedback led to a sustained decrease in task performance accuracy even after the delivery of unreliable feedback stopped – irrespective of whether reliable feedback was present (Study I) or not (Study II) in the period afterwards.

(2) **Increase in cue-congruence** – Data from Study I showed that after periods with unreliable performance feedback, perceptual inference shifted towards prior beliefs, even though the sensory information and the prior-stimulus association remained the same.

(3) **Decreased pattern distinctness in V1** – Neuroimaging data from Study II showed that unreliable feedback led to a decrease in the distinctness of multi-voxel activity patterns corresponding to visual stimuli in V1, and that this decrease was proportionate to changes in task performance.

(4) **Decreased metacognitive awareness** – Experiment 2 of Study I showed a clear decrease in metacognitive awareness during, but not after, the delivery of unreliable feedback.

(5) **A Bayesian observer model predicts the empirical results** – Data from 1000 artificial subjects, in whom the unreliable feedback delivery was simulated by the misclassification of data points in Study I, also resulted in reduced task performance and a higher cue-congruence. This paralleled the observations from the two behavioural experiments of Study I, thus supporting our hypothesis that unreliable feedback leads to the down-weighing of sensory data and the compensatory up-weighing of prior beliefs.

4.2. Novelty of results

The focus of this thesis was to understand how feedback manipulation could affect low-level perception over time. Similar to some of the previously reported studies (Herzog & Fahle, 1997; Vannucci et al., 2011; Vuvan et al., 2018; Whitson & Galinsky, 2008), we provided random feedback with chance-level validity in order to maximise uncertainty about the reliability of sensory information. In both the studies, such feedback was delivered in dedicated experimental phases called intervention runs, and their sustained effects were examined in the following phases called test runs. To control for general effects of time and the exposure to stimuli, control sessions were included in both the studies, reliable feedback replaced unreliable feedback.

In both the studies, unreliable feedback led to a sustained decrease in task performance accuracy even after the unreliable feedback stopped – irrespective of the presence (Study I) or absence (Study II) of predictive cues and reliable feedback in these periods. This sustained decrease contradicts some of the previous studies that reported improvements in performance once unreliable feedback stopped and reliable feedback was restored (Aberg & Herzog, 2012; Herzog & Fahle, 1997, 1999). However, our data are not directly comparable to these studies because of methodological differences: For instance, in Study I, the test runs, i.e., the runs that preceded and succeeded unreliable feedback, consisted of predictive cues in addition to reliable feedback – this could have

interfered with the recovery of performance that occurred in the previous studies in which a reliable feedback phase followed an unreliable feedback phase. Further, in Study II, feedback was withheld in the test runs and this could have further slowed down the recovery of task performance. Moreover, in some previous studies, difficulty levels were higher than in our experiments (e.g., 65% in Herzog & Fahle, 1997), which left more room for performance improvements and may have led to the stronger effect of reliable feedback (Liu et al., 2010, 2012).

In Study I, we further showed that unreliable feedback led to a higher reliance on cues, with an increase in the number of errors in the incorrectly predicted (i.e., incongruently cued) trials. This agrees with the Bayesian inference account of perception, according to which when the sensory data becomes less precise, the posterior – and consequently perceptual decisions – shift towards the prior beliefs (Adams et al., 2013). Similar results have also been observed in a previous study where the addition of noise to feedback shifted motor responses towards prior beliefs (Körding & Wolpert, 2004).

A simulated Bayesian observer model, which forward-modelled unreliable feedback as incorrect updating of internal likelihood distributions, replicated the effects of impaired performance and increased reliance on prior information. Here, in the simulated intervention phases, unreliable feedback was implemented as misclassifications of half of the stimuli. This led to the updating of likelihood distributions corresponding to the two stimulus categories with incorrect samples. In the simulated test phases, additional binary information was included on each trial (analogous to the cues used in the behavioural experiments of Study I). The perceptual decision-making was implemented in the simulations using a logistic regression scheme, where the weights given to priors and sensory data were re-estimated after each simulated intervention phase which had either unreliable or reliable feedback. These weights were then used to predict participants' choices. Such a scheme resulted in performance and cue-congruence changes (unreliable vs. reliable feedback) that were very similar to the behavioural data from experiments 1 and 2 in Study I.

The observed results are best supported by the explanation that sensory data gets down-weighed as a result of unreliable feedback. This process is shown schematically in Figure 4.1: The likelihood distributions corresponding to the stimuli – for example, the 45° and 135° orientations from experiment 2 of Study I – become less precise as a result

of continuous misclassification by means of unreliable feedback. This causes the likelihood estimate of the stimulus within the correct category to decrease (blue curve corresponding to 45° in Figure 4.1) and for the likelihood estimate within the incorrect category to increase (red curve corresponding to 135° in Figure 4.1), thus resulting in a higher number of erroneous decisions. In presence of priors, the brain would then attempt to compensate for the noisier sensory representations by making more use of available priors.

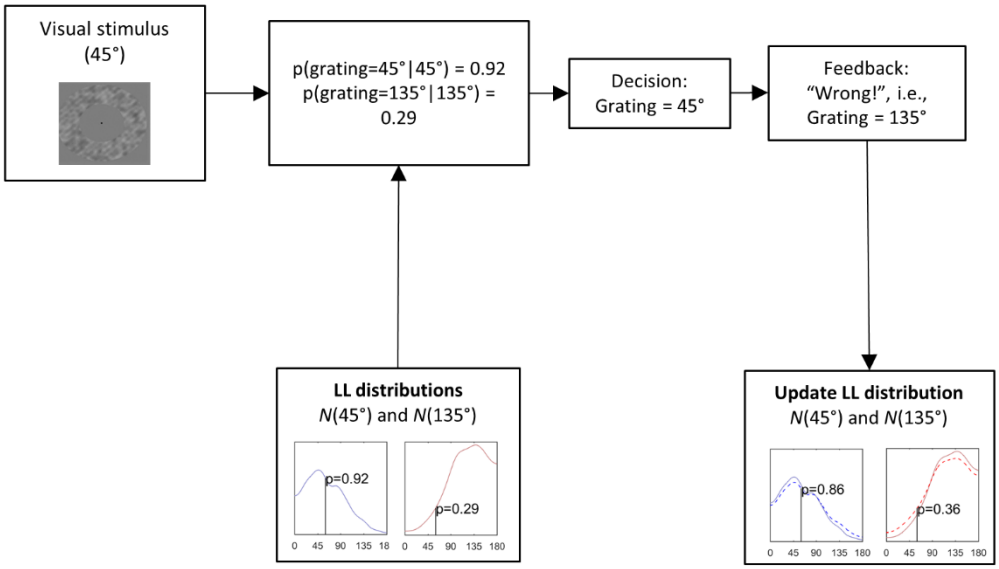

Figure 4.1: An illustration of how unreliable feedback updates the likelihood (LL) representations using an example task based on experiment 2 in Study I. The task was to identify whether the presented visual stimuli consisted of orientations 45° or 135°. Corresponding likelihood distributions are shown by the blue and red distributions (solid lines), respectively. Faulty feedback would repeatedly misclassify the representations, making them less precise over time (dotted blue and red lines) and resulting in a lower likelihood estimate for the correct stimulus category (45°, blue) and a higher likelihood estimate for the incorrect stimulus category (135°, red).

Study II further sought to understand whether the unreliable feedback would influence representations of stimuli in the sensory cortex. The primary visual cortex or V1 is most sensitive to the low-level properties of stimuli such as contrasts and spatial frequency that are critical to orientation discrimination (Avidan et al., 2002; Boynton, Demb, Glover, & Heeger, 1999; Tong et al., 2012). Hence we predicted that the precision

of sensory representations of stimuli in V1 would decrease as a result of the increase in sensory uncertainty due to unreliable feedback. Although several studies have shown the effects of sensory uncertainty on stimulus representations in sensory areas of the brain, uncertainty in these studies was induced by altering the physical properties of stimuli (Darcy et al., 2019; Hebart et al., 2012; Ludwig et al., 2016). Here, we kept the visual stimulation (i.e., stimulus orientations) constant and studied the effects of top-down beliefs about sensory uncertainty on stimulus representations. The results of Study II showed that inducing beliefs about uncertainty can indeed decrease the precision of stimulus representation in the sensory cortex.

Our analysis of the confidence data in Study I (experiment 2) revealed that confidence decreased during the delivery of unreliable feedback (intervention runs), but not afterwards (test runs). While the immediate effects are in line with a previous study (Vuvan et al., 2018), the sustained effects of unreliable feedback need to be investigated further since the analysis of session-wise slopes of the test runs indicated that there could be sustained effects of unreliable feedback over time on confidence too – while it increased over time in the reliable feedback session, it remained unchanged in the unreliable feedback session. Further, the presence of reliable feedback in the test runs in Study I could have further led to a quick recovery of confidence after each intervention run, thereby reducing the size of the sustained effects of unreliable feedback on confidence. Confidence is a well-known indicator of metacognitive awareness of one's own performance (Kleitman & Stankov, 2007; Stankov, 2000; Yeung & Summerfield, 2012) and has been shown to parallel changes in performance (Daniel & Pollmann, 2012; Guggenmos et al., 2016; Hebart et al., 2016; Vuvan et al., 2018). Our results showed that in line with performance, confidence too decreased during the delivery of unreliable feedback compared to reliable feedback. Recent studies have also highlighted another role of confidence – that of a learning signal (Daniel & Pollmann, 2012; Guggenmos et al., 2016; Hebart et al., 2016). Thus, it is possible that unreliable feedback impairs performance as a result of the decrease in the learning signal provided by confidence. To further explore this possibility, and to further understand the difference between the slopes for reliable and unreliable feedback, the effects of unreliable feedback on confidence and performance should be investigated in the absence of any feedback in the test runs.

Lastly, several non-perceptual mechanisms that could have influenced the observed behavioural and neural changes have been excluded by means of control

conditions and post-hoc analyses. The control condition with reliable feedback accounted for general learning and fatigue effects that result from performing the task for an extended period. Further, we monitored fixation actively (i.e., trials did not start unless there was fixation; see Study I) or passively (post-hoc comparison of fixation accuracies; see Study II), and this helped to exclude the possibility that eye movements led to the observed effects of unreliable feedback. Further, although there was a moderate awareness of feedback manipulation, higher awareness did not enhance our results. Similarly, motivation too remained comparable between reliable and unreliable feedback. Lastly, there was no change in the overall activity of the V1 voxels or in the performance of an independent task performed after the main experiment in Study II, thus rendering unlikely the possibility that a general decrease in attention contributed to the observed effects of unreliable feedback (Kastner et al., 1999, 1998).

4.3. Alternative accounts of the effects of unreliable feedback

As described in the previous section, our results were interpreted within the Bayesian inference framework, where we propose that stimulus distributions would get progressively more imprecise as a result of unreliable feedback on task performance. However, two other mechanisms could have alternatively occurred, which are discussed below.

First, since the delivery of unreliable feedback implies that there would be fewer trials with valid feedback compared to reliable feedback, we could observe a smaller improvement in performance. For instance, if reliable performance would lead to a 10% increase in performance, unreliable feedback would lead to a 5% improvement. Thus, although there would still be a relative decrease in performance compared to reliable feedback, the overall change would still be positive, i.e., a small performance improvement across time. However, this prediction does not consider the debilitating effect of the trials with invalid feedback, which could potentially undo the learning from the valid feedback trials. Further, for unreliable feedback to have any effect at all on behaviour, participants need to be unaware of the manipulation (or else they would simply ignore the feedback), and for this, it is important to (1) intersperse valid and invalid feedback randomly across trials within intervention runs, and (2) use a sufficiently challenging task. When using a task that satisfies these criteria, it is unlikely for participants to be able to selectively

benefit from the valid feedback trials, since it would be difficult to determine the exact trials in which feedback was valid or invalid.

A second mechanism for unreliable feedback would be one in which it provides *no* net useful information, resulting in the absence of any change in task performance. This would occur if the beneficial effects of valid feedback trials were cancelled out by the mis-learning occurring in the invalid feedback trials, since the valid and invalid trials were equal in number. This would result in the absence of any change in performance. In line with this possibility, a few previous studies that had used unreliable or random feedback have shown that it does not change the overall performance accuracy across time (Herzog & Fahle, 1997; Vannucci et al., 2011; Whitson & Galinsky, 2008).

Both of these accounts consider only trial-by-trial updating of stimulus distributions, where correctly classified trials improve the likelihood estimates of stimuli and incorrectly classified trials deteriorate the likelihood estimates. They do not take into account an important adaptive property of the brain – namely its ability to actively make inferences about the environment and its reliability (Adams et al., 2013; Behrens, Woolrich, Walton, & Rushworth, 2007; De Ridder, Vanneste, & Freeman, 2014; Schmack et al., 2016). This phenomenon has been formalised as the *active inference theory*, according to which the brain actively adapts behaviour so as to minimise sensory prediction errors (Friston, Mattout, & Kilner, 2011). The shift in responses towards priors along with the general decrease in task performance seen in our studies suggests active re-weighting by the brain in an attempt to minimise errors. It is thus possible that in an attempt to reduce the large prediction error signals continuously evoked by unreliable feedback, the brain down-weighs sensory data and up-weighs prior information, resulting in the increased prior-dependent behaviour seen in Study I.

4.4. Feedback validity and learning

The effects of varying feedback validity on perception and learning have been studied before. However, based on the research questions and the experimental designs (stimuli, task difficulty, duration etc.), the results varied vastly, ranging from deteriorations in performance to absence of improvements to improvements (Aberg & Herzog, 2012; Choi & Watanabe, 2012; Herzog & Fahle, 1997, 1999; Shibata et al., 2009; Vannucci et al., 2011;

Vuvan et al., 2018; Whitson & Galinsky, 2008). In particular, how feedback was manipulated appears to have influenced the results. For example, Shibata et al. (2009) reported an improvement in performance on providing feedback that indicated a higher performance improvement than there actually was, with the idea that the learning rate is proportionate to perceived improvement in performance. However, when providing feedback with such a bias, the overall attention and task motivation could have also been high, which could have also helped with the learning. Further, it needs to be kept in mind that perceptual learning can occur even in the absence of feedback; hence it is possible that the degree of feedback manipulation in this case was simply not sufficient to counter the naturally occurring feedback-based learning. The other study that reported performance improvement (Choi & Watanabe, 2012) provided feedback that was invalid, but still informative. In this study, the participants were trained in the task in such a way that the invalid-feedback trials suggested the presence of a previously learnt stimulus, leading to its memory retrieval every time such feedback was provided.

Studies that have used unreliable "random" feedback across all stimuli similar to our approach have reported either a decrease in performance or no change in performance (Herzog & Fahle, 1997; Vannucci et al., 2011; Vuvan et al., 2018; Whitson & Galinsky, 2008). In the study that reported no change in performance too (Herzog & Fahle, 1997), there was still a relative decrease in performance compared to reliable feedback. Further, we also measured changes in confidence and cue congruence, and all these measures indicated that unreliable feedback caused a deficit in sensory processing. Since perceptual learning is usually studied at a much slower timescale, these effects need to be studied over longer periods to understand better about the extent of these deficits and the time taken for recovery.

The studies in this thesis use the same stimuli and task to induce top-down beliefs about sensory uncertainty (i.e., deliver unreliable feedback) and to measure its effects, and based on the data from Study II, we can conclude that unreliable feedback does affect information processing in V1. However, can unreliable feedback influence perception across tasks? Two previous studies have shown evidence in favour of this – the delivery of unreliable feedback in two-choice tasks did increase pattern identification in subsequent object-identification tasks (Vannucci et al., 2011; Whitson & Galinsky, 2008). This seems compatible with real-world learning, where we seldom encounter the same sensory information all the time. In the aforementioned studies, the two tasks were

presented in the same (visual) domain and are thus likely to share some of the neural machinery. To test if unreliable feedback leads the brain to down-weigh *all* sensory data, different modalities need to be used to induce uncertainty with unreliable feedback and to measure its effects.

4.5. Was reliable feedback a good control?

The usage of reliable feedback in the control sessions instead of omitting feedback could lead to the suggestion that the observed effects of unreliable feedback were due to the *absence* of reliable feedback, rather than the *presence* of unreliable feedback. In other words, the former argument would predict that beliefs about the non-reliability of sensory data (or the volatility of the environment) were probably not induced, and we would never know this since our control condition was not "no feedback", but rather "valid feedback" – and valid feedback is known to improve performance relative to no feedback (Herzog & Fahle, 1997; Seitz et al., 2006). In spite of not having a "no feedback" condition in our studies, we find this above argument highly unlikely based on the results of behavioural performance in our studies. In our studies, performance *decreases* as a result of feedback manipulation, and previous studies which have used similar grating stimuli have already shown that learning could occur even in the absence of feedback at the at 80% performance threshold (Guggenmos et al., 2016; Liu et al., 2010, 2012). Hence, the deterioration in performance suggests that the absence of reliable feedback is not the only mechanism at play, but that a debilitating mechanism is further involved in this. Nevertheless, to understand the extent of differences between the absence of feedback and unreliable feedback, and how it relates to reliable feedback, future studies should directly compare the differences between these three conditions.

4.6. Sensory uncertainty and psychosis

Recently, theories of Bayesian learning in the brain and its potential pathological aberrations have inspired several models of psychiatric disorders such as schizophrenia (Adams et al., 2013; Corlett, Honey, Krystal, & Fletcher, 2011; Diaconescu, Hauke, & Borgwardt, 2019; Fletcher & Frith, 2009; Sterzer et al., 2018). It has been proposed that

false inferences regarding the environmental causes of sensory input data might lead to an unstable representation of the environment, as a result of which it would appear unpredictable and potentially threatening. While this notion may account for a variety of cognitive and perceptual aberrations observed in schizophrenia, it cannot easily explain one of its key features, namely, the stability of delusional beliefs, which are typically resistant to contradictory evidence. Consistent with the clinical importance of fixed delusional beliefs, it has been shown experimentally that individuals with growing delusion proneness exhibit a stronger tendency to perceive ambiguous stimuli in a manner congruent with induced prior beliefs (Schmack et al., 2013). This might engender a cycle of impaired sensory processing and compensatory strengthening of delusional beliefs, which might in turn shape perception in a belief-congruent (delusional) manner. Results from Study I demonstrate that the impairments in sensory learning induced by feedback manipulation may indeed engender the enhanced usage of prior beliefs in order to compensate for suboptimal sensory models as suggested previously (Corlett et al., 2019; Sterzer et al., 2018). Further, the results from Study II show that the induction of beliefs about reliability can manifest at the earliest levels of cortical processing, in line with a few with previous studies that have shown impairments in sensory processing in schizophrenia (Dierks et al., 1999; Gruetzner et al., 2013; Seymour et al., 2013; Silverstein, Demmin, & Bednar, 2017; Yoon et al., 2008). Thus, unreliable feedback could potentially be a useful tool to study the symptoms of aberrant perception by systematically impairing sensory precision and consequently studying its effects on perceptual inference.

4.7. Limitations

The focus of the current study was to gather a mechanistic understanding as to the effect of unreliable feedback on perception and behaviour, and one of our primary goals was to investigate the balance between sensory data and prior beliefs. To this end, we showed in Study I that behaviour shifted towards the priors (predictive cues) after phases with unreliable feedback on sensory data. In this study, the priors and unreliable feedback were presented in separate phases or runs. This separation was essential to prevent participants from strategically making decisions by attending to the cue alone and ignoring the sensory stimuli – which a participants had reported doing in a pilot version of Study I where the runs with unreliable feedback also consisted of predictive cues.

However, this separation of priors and feedback would rarely occur in actuality. Real-world scenarios might more likely resemble Körding and Wolpert (2004), where visual feedback guided motor action, and prior and feedback precision was built into this one entity, namely the visual feedback. However in this study too, the reduction in feedback precision led to prior-congruent visuo-motor learning. Thus, in spite of the controlled setting of our studies, our results mirrored the behaviour seen in visuo-motor learning when imprecise feedback was provided.

A second potential limitation could be that participants sometimes reported having detected the feedback manipulation in the debriefing questionnaire, and this could have led them to place less weight on feedback. A possible reason could be the disparity between the apparent performance indicated by the feedback and their true performance. The feedback manipulation conditions delivered positive feedback in only half of the trials (this would inevitably occur when responses are reversed in half of the trials in two-choice tasks) indicating to participants that they were performing poorly. On the other hand, their actual performance threshold was 80% (which was done to prevent a floor effect towards the end of the sessions). Future studies of unreliable feedback could better mask the external manipulation from participants, for instance by matching their apparent performance (number of trials which resulted in positive feedback) with their actual performance (actual number of correct responses) or by increasing the overall task difficulty.

In spite of several participants having indicated awareness of feedback manipulation during debriefing, there were still significant effects of unreliable feedback in both studies. Further, these ratings of awareness did not show consistent correlations with the objective measures (performance, cue-congruence, pattern distinctness etc.). This suggests that the awareness ratings themselves might have been noisy. The ratings were always collected at the end of the experiment for each participant and were thus based on their memory – which could have been noisy and often biased by the "Aha!" effect that occurs on suggestive questioning (Loftus, 2003; Topolinski & Reber, 2010). However, asking about feedback validity *during* the experiment would increase the chances of participants guessing external manipulation of feedback, which in turn would influence their perceptual decisions during the experiment. To get better indices of the awareness of feedback manipulation, future studies could include open-ended questions about feedback during debriefing which do not explicitly suggest the possibility that

feedback could have been manipulated (such as occurs when, for instance, asking for a reliability rating from participants).

Lastly, a trade-off was made in the fMRI study (Study II) between task difficulty and stimulus discriminability – the counter-clockwise and clockwise gratings with a common reference whose neural representations were compared, may have shared a large neural population. This might have reduced stimulus discriminability and consequently the size of our effects. In spite of this, we could observe the changes in pattern representations as a result of unreliable feedback. However, owing to the small sample-size and the between-group design used in this study, the results should be treated more as proof-of concept and should be replicated with larger sample sizes in the future.

4.8. Conclusions

The primary goal of this thesis was to understand how the brain would acclimate itself to an uncertain environment in which the sensory information itself does not change but rather how the environment responds to our perception of it. It is well-known that perceptual decision-making is influenced by context – for example, being given background information of a crime influences fingerprint-matching (Dror, Péron, Hind, & Charlton, 2005). Hunger and poverty have long been known to change perceptions of food and money, respectively (Bruner & Goodman, 1947; Levine, Chein, & Murphy, 1942). Several of the biases can now be successfully explained by the Bayesian perceptual inference theory. Here we used the Bayesian inference approach in combination with behavioural and neuroimaging studies to explain how perception changes in an unreliable environment. Such an environment was simulated by externally manipulating performance feedback. We hypothesised that providing such unreliable feedback to perceptual decisions would impair our reliance on sensory evidence and consequently shift the balance between prior beliefs and sensory data. Although previous studies have explored these processes to varying extents, a mechanistic understanding in relation to adaptive perceptual inference was missing. The results from our experiments revealed that in an unreliable environment, sensory data get down-weighed, leading to impaired behavioural performance and neural representations in the visual cortex, and a

compensatory shift towards prior beliefs. Further, observers seem to be aware of these behavioural changes, since a decline in subjective confidence too was observed.

Taken together, the studies in this thesis show that the induction of beliefs about the environmental reliability of sensory information systematically changes the way in which we perceive sensory information. This could be used as a framework to study how beliefs about sensory uncertainty can cause the brain to dynamically shift its balance between sensory evidence and prior beliefs. Further, unreliable feedback can be used to model neurological or psychiatric illnesses that are associated with aberrant perceptual inference. I hope that the work described in this thesis serves as a good basis for future studies that aim at studying how environmental uncertainty influences perceptual inference.

4.9. Future directions

This study is the first step towards understanding how perceptual inference adapts to unreliable feedback from the environment. We have successfully shown that Bayesian inference theories can be used to interpret this. However, several questions remain, which should be investigated in future studies:

(1) Unreliable feedback vs. absence of feedback – The decrease in task performance across time compared to baseline indicates a negative effect of unreliable feedback that it is unlikely to be merely due to the absence of feedback. However, a direct comparison between the unreliable feedback and no-feedback conditions is necessary to establish this.

(2) Long-term effects of unreliable feedback – The studies described in this thesis, together with a few previous studies, have shown that unreliable feedback leads to sustained effects even after its delivery stops. Studying the changes in performance on following days and/or weeks would help us to understand the long-term effects of unreliable feedback on learning and memory consolidation. This in turn, will give us an idea of how long it takes for the brain to re-learn the reliability of sensory information.

(3) Stimulus representation in V1 in presence of a prior – To understand the neural mechanisms underlying the increase in cue-congruence observed in Study I, an fMRI study could be designed with predictive priors in addition to sensory data. Previous studies have already shown that expectations influence stimulus representations in sensory areas (Kok et al., 2012; Schmack et al., 2013). Since unreliable feedback reduces grating stimulus precision and representational accuracy in V1, it is possible that neural stimulus representations, like behavioural choices, would become more cue-congruent following feedback manipulation.

References

Aberg, K. C., & Herzog, M. H. (2012). Different types of feedback change decision criterion and sensitivity differently in perceptual learning. *Journal of Vision*, *12*(3), 3–3. https://doi.org/10.1167/12.3.3

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in psychiatry*, *4*, 47.

Allefeld, C., & Haynes, J.-D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage*, *89*, 345–357. https://doi.org/10.1016/j.neuroimage.2013.11.043

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., & Malach, R. (2002). Contrast sensitivity in human visual areas and its relationship to object recognition. *Journal of Neurophysiology*, *87*(6), 3102–3116. https://doi.org/10.1152/jn.2002.87.6.3102

Balzan, R. P. (2016). Overconfidence in psychosis: The foundation of delusional conviction? *Cogent Psychology*, *3*(1), 1135855. https://doi.org/10.1080/23311908.2015.1135855

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Been, M., Jans, B., & De Weerd, P. (2011). Time-limited consolidation and task interference: no direct link. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(42), 14944–14951. https://doi.org/10.1523/JNEUROSCI.1046-11.2011

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Boynton, G. M., Demb, J. B., Glover, G. H., & Heeger, D. J. (1999). Neuronal basis of contrast discrimination. *Vision Research*, *39*(2), 257–269. https://doi.org/10.1016/S0042-6989(98)00113-8

Bruner, J. S., & Goodman, C. C. (1947). Value and need as organizing factors in perception. *The Journal of Abnormal and Social Psychology*, *42*(1), 33–44. https://doi.org/10.1037/h0058484

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2010). Feedback-related Negativity Codes Prediction Error but Not Behavioral Adjustment during Probabilistic Reversal Learning. *Journal of Cognitive Neuroscience*, *23*(4), 936–946. https://doi.org/10.1162/jocn.2010.21456

Choi, H., & Watanabe, T. (2012). Perceptual learning solely induced by feedback. *Vision research*, *61*, 77-82.

Corlett, P. R., Honey, G. D., Krystal, J. H., & Fletcher, P. C. (2011). Glutamatergic Model Psychoses: Prediction Error, Learning, and Inference. *Neuropsychopharmacology*, *36*(1), 294–315. https://doi.org/10.1038/npp.2010.163

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and Strong Priors. *Trends in Cognitive Sciences*, *23*(2), 114–127. https://doi.org/10.1016/j.tics.2018.12.001

Daniel, R., & Pollmann, S. (2010). Comparing the neural basis of monetary reward and cognitive feedback during information-integration category learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(1), 47–55. https://doi.org/10.1523/JNEUROSCI.2205-09.2010

Daniel, R., & Pollmann, S. (2012). Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage*, *59*(4), 3457–3467. https://doi.org/10.1016/j.neuroimage.2011.11.058

Darcy, N., Sterzer, P., & Hesselmann, G. (2019). Category-selective processing in the two visual pathways as a function of stimulus degradation by noise. *NeuroImage*, *188*, 785-793.

de Bruijn, E. R. A., Mars, R. B., & Hester, R. (2019). Processing of performance errors predicts memory formation: Enhanced feedback-related negativities for corrected versus repeated errors in an associative learning paradigm. *The European Journal of Neuroscience*. https://doi.org/10.1111/ejn.14566

De Ridder, D., Vanneste, S., & Freeman, W. (2014). The Bayesian brain: Phantom percepts resolve sensory uncertainty. *Neuroscience & Biobehavioral Reviews*, *44*, 4–15. https://doi.org/10.1016/j.neubiorev.2012.04.001

Den Ouden, H. E., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2008). A dual role for prediction error in associative learning. *Cerebral Cortex*, *19*(5), 1175–1185.

den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal Prediction Error Modulates Cortical Coupling. *Journal of Neuroscience*, *30*(9), 3210–3219. https://doi.org/10.1523/JNEUROSCI.4458-09.2010

Diaconescu, A. O., Hauke, D. J., & Borgwardt, S. (2019). Models of persecutory delusions: a mechanistic insight into the early stages of psychosis. *Molecular Psychiatry*, *24*(9), 1258–1267. https://doi.org/10.1038/s41380-019-0427-z

Dierks, T., Linden, D. E. J., Jandl, M., Formisano, E., Goebel, R., Lanfermann, H., & Singer, W. (1999). Activation of Heschl's Gyrus during Auditory Hallucinations. *Neuron*, *22*(3), 615–621. https://doi.org/10.1016/S0896-6273(00)80715-1

Dror, I. E., Péron, A. E., Hind, S.-L., & Charlton, D. (2005). When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology*, *19*(6), 799–809. https://doi.org/10.1002/acp.1130

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews. Neuroscience*, *10*(1), 48–58. https://doi.org/10.1038/nrn2536

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1), 137–160. https://doi.org/10.1007/s00422-011-0424-z

Fulvio, J. M., & Rokers, B. (2017). Use of cues in virtual reality depends on visual feedback. *Scientific Reports*, *7*(1), 16009. https://doi.org/10.1038/s41598-017-16161-3

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, *38*(12), 1861–1881. https://doi.org/10.1016/S0042-6989(97)00340-4

Gardumi, A., Ivanov, D., Hausfeld, L., Valente, G., Formisano, E., & Uludağ, K. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *NeuroImage*, *132*, 32–42. https://doi.org/10.1016/j.neuroimage.2016.02.033

Grossman, Z., & Owens, D. (2010). *An Unlucky Feeling: Overconfidence and Noisy Feedback*. Retrieved from https://escholarship.org/uc/item/13r2f3gt

Gruetzner, C., Wibral, M., Sun, L., Rivolta, D., Singer, W., Maurer, K., & Uhlhaas, P. (2013). Deficits in high- (>60 Hz) gamma-band oscillations during visual processing in

schizophrenia. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00088

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *ELife*, *5*. https://doi.org/10.7554/eLife.13388

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154. https://doi.org/10.1016/j.biopsycho.2005.04.001

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*(5), 686–691. https://doi.org/10.1038/nn1445

Hebart, M. N., Donner, T. H., & Haynes, J. D. (2012). Human visual and parietal cortex encode visual choices independent of motor plans. *Neuroimage*, *63*(3), 1393-1403.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2016). The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cerebral Cortex*, *26*(1), 118–130. https://doi.org/10.1093/cercor/bhu181

Hendriks, M. H. A., Daniels, N., Pegado, F., & Op de Beeck, H. P. (2017). The Effect of Spatial Smoothing on Representational Similarity in a Simple Motor Paradigm. *Frontiers in Neurology*, *8*, 222. https://doi.org/10.3389/fneur.2017.00222

Herzog, M. H., Aberg, K. C., Frémaux, N., Gerstner, W., & Sprekeler, H. (2012). Perceptual learning, roving and the unsupervised bias. *Vision Research*, *61*, 95–99. https://doi.org/10.1016/j.visres.2011.11.001

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research*, *37*(15), 2133–2141. https://doi.org/10.1016/S0042-6989(97)00043-6

Herzog, M. H., & Fahle, M. (1999). Effects of biased feedback on learning and deciding in a vernier discrimination task. *Vision Research*, *39*(25), 4232–4243. https://doi.org/10.1016/S0042-6989(99)00138-8

Hohwy, J. (2012). Attention and Conscious Perception in the Hypothesis Testing Brain. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00096

Hol, K., & Treue, S. (2001). Different populations of neurons contribute to the detection and discrimination of visual motion. *Vision Research*, *41*(6), 685–689. https://doi.org/10.1016/S0042-6989(00)00314-X

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E. M., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*(2), 519–530. https://doi.org/10.1016/j.neuron.2013.09.009

Jiang, J., Summerfield, C., & Egner, T. (2013). Attention Sharpens the Distinction between Expected and Unexpected Percepts in the Visual Brain. *Journal of Neuroscience*, *33*(47), 18438–18447. https://doi.org/10.1523/JNEUROSCI.3308-13.2013

Kahnt, T., Grueschow, M., Speck, O., & Haynes, J.-D. (2011). Perceptual Learning and Decision-Making in Human Medial Frontal Cortex. *Neuron*, *70*(3), 549–559. https://doi.org/10.1016/j.neuron.2011.02.054

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140169.

Karvelis, P., Seitz, A. R., Lawrie, S. M., & Seriès, P. (2018). Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. *ELife*, *7*. https://doi.org/10.7554/eLife.34115

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased Activity in Human Visual Cortex during Directed Attention in the Absence of Visual Stimulation. *Neuron*, *22*(4), 751–761. https://doi.org/10.1016/S0896-6273(00)80734-5

Kastner, S., Weerd, P. D., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of Directed Attention in the Human Extrastriate Cortex as Revealed by Functional MRI. *Science*, *282*(5386), 108–111. https://doi.org/10.1126/science.282.5386.108

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161–173. https://doi.org/10.1016/j.lindif.2007.03.004

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kok, P., Brouwer, G. J., Gerven, M. A. J. van, & Lange, F. P. de. (2013). Prior Expectations Bias Sensory Representations in Visual Cortex. *Journal of Neuroscience*, *33*(41), 16275–16284. https://doi.org/10.1523/JNEUROSCI.0742-13.2013

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265-270.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244.

Köther, U., Veckenstedt, R., Vitzthum, F., Roesch-Ely, D., Pfueller, U., Scheu, F., & Moritz, S. (2012). "Don't give me that look" — Overconfidence in false mental state perception in schizophrenia. *Psychiatry Research*, *196*(1), 1–8. https://doi.org/10.1016/j.psychres.2012.03.004

Lempert, K. M., & Tricomi, E. (2015). The Value of Being Wrong: Intermittent Feedback Delivery Alters the Striatal Response to Negative Feedback. *Journal of Cognitive Neuroscience*, *28*(2), 261–274. https://doi.org/10.1162/jocn_a_00892

Levine, R., Chein, I., & Murphy, G. (1942). The relation of the intensity of a need to the amount of perceptual distortion: A preliminary report. *Journal of Psychology; Provincetown, Mass., Etc.*, *14*, 283–293.

Liu, J., Lu, Z.-L., & Dosher, B. A. (2010). Augmented Hebbian reweighting: interactions between feedback and training accuracy in perceptual learning. *Journal of Vision*, *10*(10), 29. https://doi.org/10.1167/10.10.29

Liu, J., Lu, Z.-L., & Dosher, B. A. (2012). Mixed training at high and low accuracy levels leads to perceptual learning without feedback. *Vision Research*, *61*, 15–24. https://doi.org/10.1016/j.visres.2011.12.002

Loftus, E. F. (2003). Make-believe memories. *American Psychologist*, *58*(11), 867–873. https://doi.org/10.1037/0003-066X.58.11.867

Ludwig, K., Sterzer, P., Kathmann, N., & Hesselmann, G. (2016). Differential modulation of visual object processing in dorsal and ventral stream by stimulus visibility. *Cortex*, *83*, 113-123.

Miltner, W. H., Braun, C. H., & Coles, M. G. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a 'generic' neural system for error detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. https://doi.org/10.1162/jocn.1997.9.6.788

Misaki, M., Luh, W.-M., & Bandettini, P. A. (2013). The effect of spatial smoothing on fMRI decoding of columnar-level organization with linear support vector machine. *Journal of Neuroscience Methods*, *212*(2), 355–361. https://doi.org/10.1016/j.jneumeth.2012.11.004

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784. https://doi.org/10.1126/science.4023713

Moritz, S., Ramdani, N., Klass, H., Andreou, C., Jungclaussen, D., Eifler, S., … Zink, M. (2014). Overconfidence in incorrect perceptual judgments in patients with schizophrenia. *Schizophrenia Research: Cognition*, *1*(4), 165–170. https://doi.org/10.1016/j.scog.2014.09.003

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943–1948. https://doi.org/10.1016/j.neuroimage.2009.02.047

O'Reilly, J. X., Jbabdi, S., & Behrens, T. E. J. (2012). How can a Bayesian approach inform neuroscience? *The European Journal of Neuroscience*, *35*(7), 1169–1179. https://doi.org/10.1111/j.1460-9568.2012.08010.x

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98. https://doi.org/10.1038/nn802

Parr, T., Rees, G., & Friston, K. J. (2018). Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00061

Paulus, M. P., & Yu, A. J. (2012). Emotion and decision-making: affect-driven belief systems in anxiety and depression. *Trends in Cognitive Sciences*, *16*(9), 476–483. https://doi.org/10.1016/j.tics.2012.07.009

Peirce, C. S., & Jastrow, J. (1884). On Small Differences in Sensation. *Memoirs of the National Academy of Sciences*, *3*, 75–83.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2006). Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Research*, *46*(19), 3177–3197. https://doi.org/10.1016/j.visres.2006.03.022

Pfabigan, D. M., Alexopoulos, J., Bauer, H., & Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. *Psychophysiology*, *48*(5), 656–664. https://doi.org/10.1111/j.1469-8986.2010.01136.x

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(21), 6059–6064. https://doi.org/10.1073/pnas.1522551113

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–381. https://doi.org/10.1038/26475

Sagi, D., & Julesz, B. (1984). Detection versus Discrimination of Visual Orientation. *Perception*, *13*(5), 619–628. https://doi.org/10.1068/p130619

Schmack, K., Gòmez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., … Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(34), 13701–13712. https://doi.org/10.1523/JNEUROSCI.1778-13.2013

Schmack, K., Weilnhammer, V., Heinzle, J., Stephan, K. E., & Sterzer, P. (2016). Learning What to See in a Changing World. *Frontiers in Human Neuroscience*, *10*, 263. https://doi.org/10.3389/fnhum.2016.00263

Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, *5*(7), 532. https://doi.org/10.1038/nrn1427

Seitz, A. R., Nanez, J. E., Holloway, S., Tsushima, Y., & Watanabe, T. (2006). Two cases requiring external reinforcement in perceptual learning. *Journal of Vision*, *6*(9), 9–9. https://doi.org/10.1167/6.9.9

Seymour, K., Stein, T., Sanders, L. L. O., Guggenmos, M., Theophil, I., & Sterzer, P. (2013). Altered Contextual Modulation of Primary Visual Cortex Responses in Schizophrenia. *Neuropsychopharmacology*, *38*(13), 2607–2612. https://doi.org/10.1038/npp.2013.168

Shibata, K., Yamagishi, N., Ishii, S., & Kawato, M. (2009). Boosting perceptual learning by fake feedback. *Vision research*, *49*(21), 2574-2585.

Silverstein, S. M., Demmin, D. L., & Bednar, J. A. (2017). Computational Modeling of Contrast Sensitivity and Orientation Tuning in First-Episode and Chronic Schizophrenia. *Computational Psychiatry*, *1*, 102–131. https://doi.org/10.1162/CPSY_a_00005

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence*, *28*(2), 121–143. https://doi.org/10.1016/S0160-2896(99)00033-1

Stefanics, G., Stephan, K. E., & Heinzle, J. (2019). Feature-specific prediction errors for visual mismatch. *NeuroImage*, *196*, 142–151. https://doi.org/10.1016/j.neuroimage.2019.04.020

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., … Corlett, P. R. (2018). The Predictive Coding Account of Psychosis. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2018.05.015

Thorndike, E. L. (1913). *The psychology of learning*. Teachers College, Columbia University.

Tong, F., Harrison, S. A., Dewey, J. A., & Kamitani, Y. (2012). Relationship between BOLD amplitude and pattern classification of orientation-selective activity in the human visual cortex. *NeuroImage*, *63*(3), 1212–1222. https://doi.org/10.1016/j.neuroimage.2012.08.005

Topolinski, S., & Reber, R. (2010). Gaining Insight Into the "Aha" Experience. *Current Directions in Psychological Science*, *19*(6), 402–405. https://doi.org/10.1177/0963721410388803

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579. https://doi.org/10.1038/21176

Ullsperger, M., Fischer, A. G., Nigbur, R., & Endrass, T. (2014). Neural mechanisms and temporal dynamics of performance monitoring. *Trends in Cognitive Sciences*, *18*(5), 259–267. https://doi.org/10.1016/j.tics.2014.02.009

van Vugt, F. T., & Tillmann, B. (2015). Auditory feedback in error-based learning of motor regularity. *Brain Research*, *1606*, 54–67. https://doi.org/10.1016/j.brainres.2015.02.026

Vannucci, M., Mazzoni, G., & Cartocci, G. (2011). Lack of control enhances accurate and inaccurate identification responses to degraded visual objects. *Psychonomic Bulletin & Review*, *18*(3), 524–530. https://doi.org/10.3758/s13423-011-0083-z

Vickers, D. (2014). *Decision processes in visual perception*. Academic Press.

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology: CB*, *22*(18), 1641–1648. https://doi.org/10.1016/j.cub.2012.07.010

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik* (Vol. 9). Voss.

Vuvan, D. T., Zendel, B. R., & Peretz, I. (2018). Random Feedback Makes Listeners Tone-Deaf. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-25518-1

Whitson, J. A., & Galinsky, A. D. (2008). Lacking Control Increases Illusory Pattern Perception. *Science*, *322*(5898), 115–117. https://doi.org/10.1126/science.1159845

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Yoon, J. H., Tamir, D., Minzenberg, M. J., Ragland, J. D., Ursu, S., & Carter, C. S. (2008). Multivariate Pattern Analysis of Functional Magnetic Resonance Imaging Data Reveals Deficits in Distributed Representations in Schizophrenia. *Biological Psychiatry*, *64*(12), 1035–1041. https://doi.org/10.1016/j.biopsych.2008.07.025

Acknowledgements

This doctoral thesis would not have materialised without the support of my thesis supervisor Prof. Dr. Philipp Sterzer. Philipp showed immense trust in a nascent research idea and gave me the right framework to convert it to a full-blown PhD thesis over the past years. Philipp's excellent guidance was not only crucial to the thesis, but also to my growth as a researcher and a person.

I would also like to thank the members of the Visual perception Lab at Charité. First of all, I am deeply indebted to PD Dr. Matthias Guggenmos – I couldn't have done this without his support and encouragement all through the past 4 years, both related to work and outside it. I am also grateful to him for his suggestions on the dissertation and for help with translating the abstract. Next, I would like to sincerely thank Dr. Heiner Stuke and Dr. Marcus Rothkirch for their timely help especially with data acquisition, analysis and paper-writing. I am thankful to all of my current and previous lab mates for their critical suggestions at lab presentations, for being excellent pilot participants, and for the long Mensa conversations and dinners.

I would also like to thank my PhD co-supervisor Prof. Dr. John-Dylan Haynes as well for his valuable guidance during the progress report meetings and for having been my official supervisor at HU Berlin.

Also indispensable to the thesis work was the support from the Bernstein Centre for Computational Neuroscience (BCCN) Berlin and the graduate programme "Sensory Computation and Neural Systems" GRK 1589/2. The GRK provided me with a scholarship for 3.5 years in addition to financially supporting a large part of my thesis work and conference attendances. In particular, I would like to thank Dr. Robert Martin, Margret Franke and Camilla Groiss who helped me glide through organisational hassles, especially in light of my limited German. I also really liked the overall hospitable environment at BCCN during communal events such as the PhD symposia, colloquia and retreats, and enjoyed the time with my BCCN cohorts. I would also like to thank Jana Lahmer from the Life Sciences faculty at HU Berlin for timely replies to innumerable questions about thesis submission.

I am also eternally grateful to my incredible friends, both from Berlin and outside. They helped me through some really tough days.

Lastly, I am thankful to my beloved Ullas for staying by my side and for bringing colour to my life outside work. I am also grateful to our families for their support and words of wisdom.

Declaration of Independent Work

I state that the following are true:

- that no collaboration with commercial doctoral degree supervisors took place;

- that I acknowledge the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences at HU Berlin, as amended on 5 March 2015 ;

- that the doctoral thesis, or parts of it, have not already been submitted to, or approved or rejected by, another academic institution;

- that I have not applied for a doctoral degree elsewhere or have a corresponding doctoral degree;

- that the doctoral thesis was written by me independently based on the stated resources and aids, under the provisions of § 6 (3), and that contributions from co-authors have been explicitly mentioned at the beginning of the thesis;

- that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by;